KALIE: Fine-Tuning Vision-Language Models for Open-World Manipulation without Robot Data

Anonymous Author(s)
Affiliation
Address
email

Abstract: Building generalist robotic systems involves effectively endowing robots the capabilities to handle novel objects in an open-world setting. Inspired by the advances of large pre-trained models, we propose Keypoint Affordance Learning from Imagined Environments (KALIE), which adapts pre-trained Vision Language Models (VLMs) for robotic control in a scalable manner. Instead of directly producings motor commands, KALIE controls the robot by predicting point-based affordance representations based on natural language instructions and visual observations of the scene. The VLM is trained on 2D images with affordances labeled by humans, bypassing the need for training data collected on robotic systems. Through an affordance-aware data synthesis pipeline, KALIE automatically creates massive high-quality training data based on limited example data manually collected by humans. We demonstrate that KALIE can learn to robustly solve new manipulation tasks with unseen objects given only 50 example data points. Compared to baselines using pre-trained VLMs, the our approach consistently achieves superior performances. [1]

Keywords: Vision-Language Model, Data Synthesis, Fine-Tuning, Manipulation

1 Introduction

The capability to handle an open set of objects, behaviors, and task specifications is essential to the development of generalist robotic systems. Existing learning methods for robotic control can require extensive amounts of data collected on embodied systems [1, 2, 3]. The diversity and quality of the collected data determine the generalization capability that these methods can achieve, which is subject to robotics expertise and manual labors that humans can provide. How can we endow robots with generalizable skills for solving an open set of tasks in a scalable manner?

Large pre-trained models offer promising tools with their generalist visual understanding and commonsense reasoning abilities [4, 5, 6, 7]. Prior works have shown that pre-trained Large Language Models (LLMs) and Vision Language Models (VLMs) can be directly applied to robotics control through prompt engineering in a zero-shot manner [8, 9, 10, 11]. However, despite the capability of generalizing to unseen tasks, such systems often suffer from instability and require significant hard-coded domain knowledge to compensate the pre-trained models' limited knowledge about robotic control. While fine-tuning on robotic data can mitigate this issue and proves more sample-efficient than training policies from scratch [12, 13], the largest available datasets [14, 15, 16] for robotic control is still far from being comparable to the Internet-scale data used for pre-training the large models with billions of parameters. It remains a grand challenge to effectively employ and adapt pre-trained large models for robotic control.

In this paper, we aim to study an alternative solution to this challenge by training large models for manipulation without data collected with on robots. Our key insight is to use visual affordances to

[1] Project webpage: https://sites.google.com/view/kalie-vlm

Submitted to the 8th Conference on Robot Learning (CoRL 2024). Do not distribute.

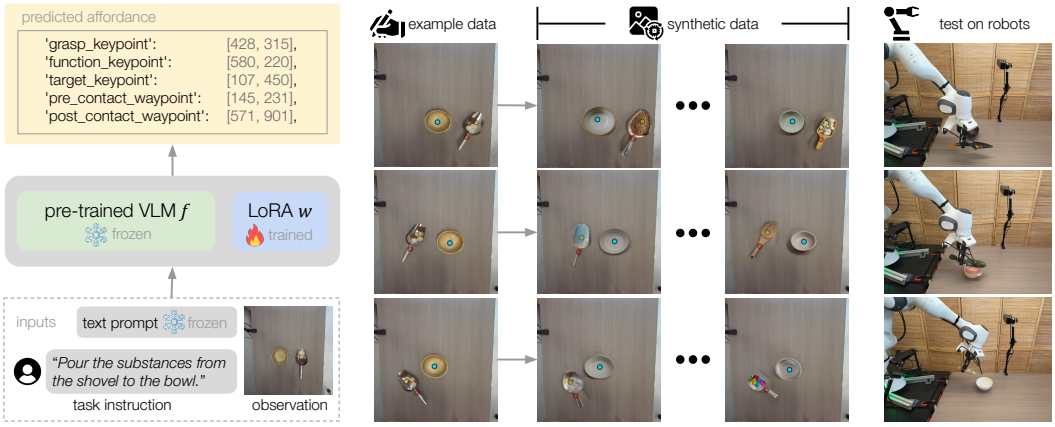

Figure 1: **Overview of KALIE**. By fine-tuning a pre-trained VLM, KALIE predicts the point-based affordance representation given the input task instruction and visual observation. Based on limited example data collected by humans, KALIE generates synthetic data with high diversity while preserving the task semantics and the keypoint annotations. The fine-tuned VLM can robustly generate motions for manipulation tasks with unseen objects and scene arrangements.

guide robotic control and leverage the broad knowledge incorporated in large pre-trained models to efficiently learn to predict the affordances. Built upon point-based affordance representations defined on 2D images from Manuelli et al. [17] and Fang et al. [10], we fine-tune a VLM on labeled affordance data. Humans can easily collect such affordance data by randomizing scenes for a target task (e.g., cleaning a table), taking images through the camera, and annotating the affordances on the image, which bypasses the need for collecting demonstration trajectories through teleoperation the robots or hard-coded policies. The main challenges are 1) how to repurpose the VLM pre-trained for visual-question answering for efficient affordance learning, and 2) how to efficiently utilize human supervisions to create training data that can cover diverse scenarios.

To this end, we propose Keypoint Affordance Learning from Imagined Environments (KALIE). As shown in Fig. 1, our approach fine-tunes a pre-trained VLM to the point-based affordance representation given the input task instruction and visual observation. Starting with limited example data of manually arranged scenes and annotated affordance labels, KALIE automatically creates massive and diverse synthetic images to scale up the training data. To stay faithful to the task semantics and keypoint annotations while diversifying the data distribution as much as possible, we propose an affordance-aware data synthesis pipeline using a pre-trained diffusion model [18, 19] with additional contexts. We evaluate KALIE on various manipulation tasks involving tool-use, articulated objects, and deformable objects. KALIE consistently solves target tasks with diverse unseen objects and initial arrangements and achieves superior performances compared to baselines using pre-trained VLMs [9, 10].

2 Keypoint Affordance Learning from Imagined Environments

We propose Keypoint Affordance Learning from Imagined Environments (KALIE) to adapt pre-trained Vision Language Models (VLMs) to acquire generalizable skills without robot experiences. In this section, we will first define the affordance prediction problem in the few-shot setting using point-based affordances labeled by human experts. Next, we will introduce a novel affordance-aware data synthesis recipe to diversify the training data, which automatically generates massive high-quality data based on the example data collected only for limited scenarios. Then, we will describe our VLM fine-tuning approach and discuss the key design options. Lastly, we summarize the overall system at the end of this section.

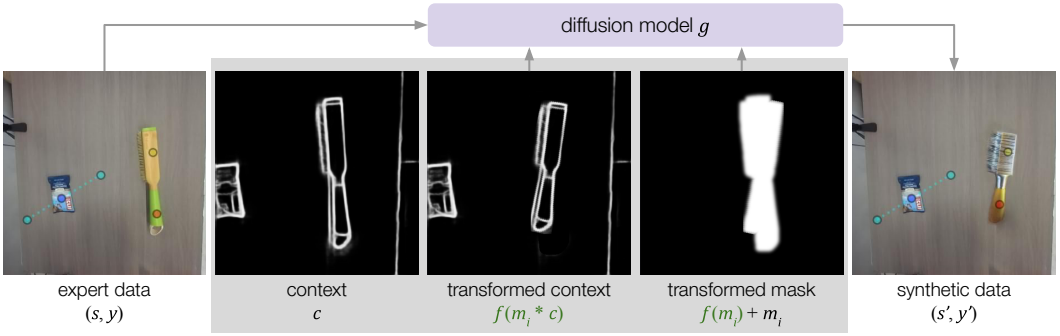

Figure 2: **Affordance-aware data synthesis.** KALIE employs the inpainting capability of a diffusion model to generate synthetic data. To diversify the scenes while staying faithful to the task semantics and the keypoint annotations, KALIE computes and transforms the context such as soft edges as additional inputs guiding the diffusion model.

2.1 Problem Statement

We consider the problem of open-world robotic manipulation involving unseen objects and initial arrangements of the scenes. As shown in Fig. 1, each task is single-stage and is specified by a free-form language description l, such as "*Using the brush to sweep the snack package.*" The robot observes an RGBD image from a third-person camera and performs a 6-DoF motion trajectory in open loop to complete the task.

To tackle this problem, we employ a VLM f pre-trained on Internet-scale data [6]. Following the practice of [10], we query the VLM to produce point-based affordance representations to guide a motion planner to generate motions. The VLM takes as inputs the prompt ρ, the task instruction l, and the input image s and predict the affordance representation y as:

$$\hat{y} = f(\rho, l, s), \tag{1}$$

where y contains a set of keypoints, including the grasp keypoint, the function keypoint, the target keypoint, the pre-contact waypoint, and the post-contact waypoint, specified by 2D coordinates on the image (see Fig. 1). Additional properties such as the height and the orientation of the gripper will also be decided for each task. Based on the affordance representation, a low-level motion generator computes a motion trajectory to complete the task. We assume the desired motion to solve each task can be specified by the same subset of these points (e.g., the sweeping with a brush requires all five points, and drawer closing requires everything but the grasp point, as shown in Section 3.3), but the specific coordinates of these points depend on the objects and their poses with respect to the robot. More detail about definition of the affordance representation and the prompt design will be elaborated in the Appendix.

In this work, we consider tasks and objects which are challenging for VLMs to handle in zero-shot. In contrast to [10], we consider a few-shot learning setting, in which we fine-tune the pre-trained VLM to acquire and improve skills for unseen scenarios on limited and non-robot data. Notably, the point-based representations enable us to outsource motion generation to the low-level motion planner and focus on only the affordance prediction problem. Therefore, training only requires data of pairs of observed image s and the ground truth keypoints y collected by human experts or generated automatically as described below. We assume access to an example dataset $\mathcal{D}$ containing a limited number of (s, y) pairs for each target task. In our experiments, we assume the most extreme case, in which the data is collected on *a single set of* objects for each task and the fine-tuned model is evaluated on unseen object sets.

2.2 Affordance-Aware Data Synthesis

To scale up the training of the VLM to enhance its generalization capabilities to unseen scenarios, we automatically synthesize a training dataset $\mathcal{D}'$ to cover a wide range of environments. Each of

the new data points $(s', y') \in \mathcal{D}'$ is synthesized by modifying an existing data point $(s, y) \in \mathcal{D}$ collected by human experts.

Following the practice of [20], we compute the segmentation mask of the object in the scene using open-vocabulary segmentation [21] and then inpaint the masked region with a diffusion model [18]. However, naively inpainting the masked region can lead to undesired results. Without a direct mechanism to specify the geometric properties of the object, it would be hard to diversify the inpainted images in a way that can cover the desired distribution of testing scenarios. Moreover, there is usually a discrepancy between the appearance of the inpainted object and the original keypoint annotation, introducing the need of manually re-labeling keypoints. To generate massive, high-quality data without additional manual labor, we need to ensure that the generated images stay faithful to the context of the target task and the annotated keypoints, while aggressively diversifying the environments as much as possible.

To tackle this challenge, we design an affordance-aware data synthesis pipeline by leveraging additional context images to guide the generation process, as shown in Fig. 2. Specifically, we employ the ControlNet [19] diffusion model g, which takes inputs as the input image s, the segmentation mask m, a context image c, and a language description of the object o and generates the new image as $s' \sim g(\cdot|s, m, c, o)$. We would like c to be a compact representation of the object's geometric properties that provide clues about the affordances, while minimizing other detailed visual information to leave enough free reign for the diffusion model. By inpainting an image s' in accordance with c, we hope to obtain new objects of the same point-based affordances. In this work, we choose to use a soft edge map as c computed by an external image processing algorithm [22], which outlines the contours and parts of the object.

To cover testing objects of unseen shapes, we introduce additional randomization operations to the geometry of the synthesized object. Directly modifying either the object's appearance in the pixel space can easily affect other parts of the image and create artifacts. Instead, we propose to apply transformation $h(\cdot)$ on the compact context c before calling the diffusion model G to inpaint the image. The transformation function $h(\cdot)$ can include basic data augmentation operations such as random scaling, translation, and rotation, as well as additional operations such as elastic distortion. To modify the context c, we transform the region under the masked as $h(m * c)$. The region to be inpaint on the original image now becomes $m + h(m)$, which includes the transformed silhouette of the object $h(m)$ in addition to the white space left by removing the original object m. Accordingly, we also apply the same transformation to the annotated keypoints y with the context image to keep them consistent. Therefore, the sampling process with the transformation can be written as:

$$\begin{aligned} s' &\sim g(\cdot|s, h(m) + m, h(m * c), o) \\ y' &= h(y), \end{aligned} \tag{2}$$

where we slightly overload the notation by using $h(\cdot)$ to denote the transformation applied to both the images and the keypoint coordinates.

2.3 Efficient Adaptation for Keypoint Affordance Prediction

We fine-tune the VLM f to predict point-based affordance representations. To adapt f, which is pre-trained for visual-question answering by predicting tokens, we need to make our design choices around two considerations. First, how to represent the point-based affordances so that we can effectively re-purpose or modify the VLM's prediction head for affordance prediction. Second, how to perform sample-efficient fine-tuning to utilize the pre-trained VLM's pre-trained capabilities while endowing it with the additional knowledge from the new dataset $\mathcal{D}$.

We investigate two design options for the prediction head: 1) **Regression Head** [23] adds an additional linear layer on top of the last hidden state of the VLM to directly predict the $x - y$ coordinates of the keypoint affordances. In input to the VLM is the image along with appropriate task instructions and the last hidden state of the last token is used. 2) **Natural Language Affordance Prediction** fine-tunes the VLM to output a well-formatted natural language that includes text-based keypoint

affordances with an example shown in Fig. 1. Each keypoint affordance is represented by the $x - y$ coordinates as integers normalized between a predefined range.

These design options resort to the pre-trained weights and the new dataset in different manners. Empirically as shown in Fig. 3, we found those two design choices achieve similar performances. As Natural Language Affordance Prediction aligns more closely with other applications of VLMs such as Visual Question Answering, we choose to use this design option as our main method.During training, we convert the ground truth affordance label y into the corresponding format to compute the losses. For both design options, we use Low-Rank Adaptation (LoRA) [24] to fine-tune the VLM. We use a L2 regression loss for Regression Head and the cross entropy loss for Natural Language Affordance Prediction.

2.4 System Summary

The overall pipeline of KALIE is summarized in Algorithm 1. Starting with the original example dataset $\mathcal{D}$, KALIE automatically create $\mathcal{D}'$ to fine-tune the VLM f. The data synthesis pipeline operates in an object-centric manner, iteratively process and inpaint each of the M objects in the scene as explained in Section 2.2, using each object's segmentation mask m_i and description o_i. During this process, a VLM (which can be the same with f or a different model) is used to sample an alternative description of the object to guide the inpainting process of the diffusion model g. The generated data is combined with the example dataset to adapt f by optimizing the weights w as explained in Section 2.3. More implementation details will be explained in the Appendix.

Algorithm 1 Keypoint Affordance Learning from Imagined Environments (KALIE)

Inputs: Pre-trained VLM f, task instruction l, example dataset D, desired dataset size N.

$D' \leftarrow \varnothing$
while $|D'| < N$ **do**
  Sample (s, y) from D.
  Compute the context image c given s.
  Extract the object descriptions $\{o_i\}_{i=1}^{M}$ for each object i given s and l.
  $s'_0 \leftarrow s$
  for $i = 1, ..., M$ **do** (Section 2.2)
    Compute the segmentation mask m_i for the i-th object given s and o_i.
    Sample a new description o'_i given o_i.
    Transform the mask, the context as $h(m_i)$ and $h(m_i * c)$.
    Transform the keypoints y_i on the i-th object as $h(y_i)$.
    Inpaint the image as $s'_i \sim G(\cdot|s'_{i-1}, h(m_i) + m_i, h(m_i * c), o'_i)$.
  end for
  Merge the transformed keypoints into y'.
  if s'_M pass the filter: **then**
    $D' \leftarrow D' \cup \{(s'_M, y')\}$
  end if
end while
Fine-tune f on $D \cup D'$ by optimizing the weights w (Section 2.3).

3 Experiments

We design our experiments to investigate the following questions: 1) Can KALIE synthesize diverse and high-quality data for affordance prediction? 2) Can KALIE fine-tune pre-trained VLMs to improve their performances on challenging manipulation tasks with unseen objects?

3.1 Experimental Setup

Environments. Our experiments are conducted in a real-world table-top manipulation environment with a 7-DoF robot arm and a parallel jaw gripper. A top-down RGBD camera is used to receive visual observation of the environment.

Tasks. We design various table-top manipulation tasks with a diverse set of daily objects

Training and evaluation protocols. We fine-tune CogVLM [6], an open-source VLM of 17 billion parameters. To fine-tune the VLM, 50 input images are collected by humans through randomly setting up the scenes based on the context of the task and annotating the affordance representations through a graphical user interface. Though data synthesis with diffusion models, we generate an additional set of 500 pairs of images and keypoints. For each target task, test the fine-tuned model on 15 trials with three sets of unseen objects.

Baselines. We compare KALIE with two baselines that employ VLMs for robotic manipulation, **VoxPoser** [9] and **MOKA** [10]. In contrast to KALIE , which pre-trains a much smaller open-source VLM, the two baselines use the OpenAI GPT-4V[4]. We use the collected example data as the in-context examples in MOKA.

Methods	Table Sweeping	Drawer Closing	Towel Hanging	Trowel Pouring	USB Unplugging
VoxPoser [9]	3/15	8/15	1/15	0/15	0/15
MOKA [10]	9/15	9/15	5/15	7/15	2/15
KALIE (Ours)	**13**/15	**15**/15	**12**/15	**11**/15	**9**/15

Table 1: **Task success rates.** We evaluate KALIE on five manipulation tasks involving tool use, deformable objects, and articulated objects. KALIE robustly solve these tasks and consistently achieves superior performances compared to baselines.

3.2 Quantitative Results

Comparative results. We evaluate KALIE across 5 manipulation tasks involving tool objects, articulated objects, and deformable objects. Compared to the baseline methods using pre-trained VLMs, KALIE consistently achisves higher success rates. Baseline methods are often subject to critical failures including perception errors, incorrect output formats, and unsuitable predicted affordances. The VLM fine-tuned by KALIE demonstrate to robustly solve the tasks without these issues while most failures are due to inaccurate point predictions.

Ablative study. In Fig. 3, we present some ablation experiment results examining the effectiveness of Imagined Environments and our choice of Natural Language Affordance Prediction against Regression Head. We report the test Mean Square Error (MSE) on a held-out set with unseen objects and ground-truth annotations. The coordinates are normalized between 0 and 999 with respect to the actual height and width of the image. The train set consists of 50 real images collected and 500 synthesized images, and the test set consists of 50 real images of novel objects excluded from the train set. We observe that Imagined Environments significantly improve the test MSE and ours with Natural Language Affordance Prediction works comparably with Regression Head, while staying consistent with the general natural language interface of VLMs for vision-language tasks.

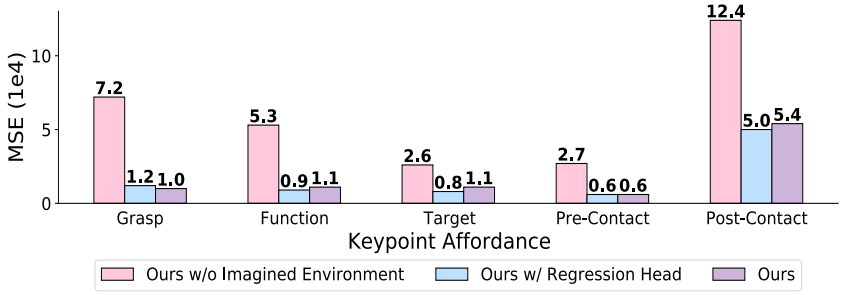

Figure 3: **Ablative study.** The ablation experiments are carried out on the sweeping task. Mean Square Error (MSE) for each keypoint affordance on the a test set of novel objects is reported.

3.3 Qualitative Results

In Fig. 4, we show examples of the synthetic images. Each row contains the data for a task. The first row is the example data and the remaining rows are the synthetic data. KALIE demonstrates to generate diverse samples while preserving the task context and the original affordance lables.

We show examples of task execution in Fig. 5. The VLM fine-tuned by KALIE successfully predict the affordance and solve each of the tasks.

Figure 4: **Data synthesis examples.** We show the example data and synthetic data generated by KALIE for each of the five tasks. Each row contains the data for a task. The first row is the example data and the remaining rows are the synthetic data.

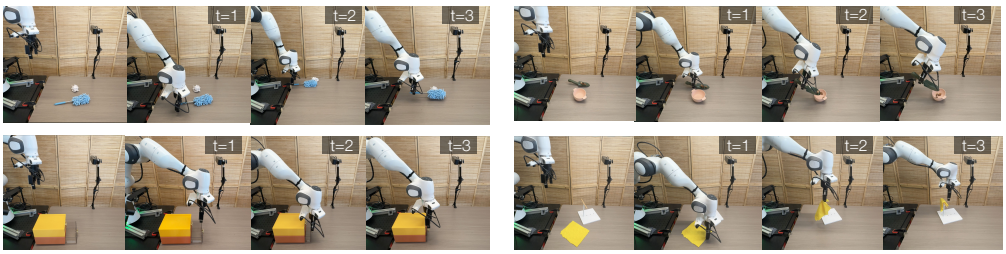

Figure 5: **Affordance prediction and task execution.** We show examples of the predicted affordance representations and the key frames of task execution. Across various manipulation tasks, KALIE demonstrates to robustly solve the task.

4 Related Work

Adaptation of vision language models. Due to their versatile nature, Vision Language Models (VLMs) can be effectively fine-tuned to accommodate a variety of downstream tasks [25, 26, 27, 28].

Specifically, to facilitate the prediction of spatially grounded outputs, previous research has investigated methods such as sets of marks [29], scaffolding [4], and coordinate-based bounding boxes [30, 31, 6]. Of these methods, coordinate-based references are particularly adaptable, capable of pinpointing any location within an image. Thus, we employ this approach for predicting keypoints. Our contribution does not lie in proposing a new method for fine-tuning VLMs, but rather in integrating and adapting VLM fine-tuning for robotic control.

Robotic control with large models. An increasing number of works have employed foundation models in robotics through pre-training or fine-tuning on robot data manually collected through teleoperation or scripted policies [12, 32, 13, 33]. Due to the lack of datasets that can cover the vast complexity and diversity of robotics applications, most of the approaches focus on specific task categories such as grasping and object rearrangements. The generalization capability of the trained model is also constrained to the distribution of objects and environments covered by the manually collected datasets. Alternatively, other works have attempted to combine and prompt pre-trained models to solve unseen tasks in zero-shot manners [8, 9, 10, 11, 34]. However, the performance of these works is usually subject to the capability of the pre-trained models, as well as non-trivial expert knowledge and manual labor to design prompts and in-context examples. In contrast to these approaches, the proposed method fine-tunes a VLM to robustly solve the target tasks. To avoid the need of collecting extensive amount of robot data, our model train the VLM to predict the point-based affordance representation from [10] and employ a diffusion model to automatically synthesize massive, high-quality data.

Data synthesis in robotics. To alleviate the data bottleneck in robot learning, prior work has investigated various approaches to augment and synthesize data. Data augmentation, especially random image transformations, have been widely used to improve generalization to unseen visual inputs [35, 36]. These random operations can effectively improve the models generalization capabilities to unseen visual inputs during the test time, but cannot extend the model's capabilities beyond the coverage of the training distribution. Domain randomization has also been broadly used to train robust models for robotic control [37, 38, 39, 40] when collecting simulated robot experiences. In contrast, our method directly diversify the training data without the need for a physical simulator. Recent works in computer vision and robotics have leveraged deep generative models, such as diffusion models, to synthesize unseen environments by leveraging broad knowledge learned from Internet-scale images [41, 42, 43, 44, 20, 45, 46, 47, 48]. While the prior work can easily generate defected samples, we propose an affordance-aware data synthesis approach that enable the diffusion model to generate diverse data with much higher quality and consistent affordance annotations.

5 Conclusion

In this work, we propose KALIE to fine-tune pre-trained VLMs to predict affordances for robotic manipulation with open sets of objects and initial arrangements of scenes. Using the proposed affordance-aware data synthesis pipeline, we generate massive high-quality data to scale up training without extensive manual labors or domain expertise with robots. We fine-tune the VLM on the generated data to predict the point-based affordance representations. Across various manipulation tasks involving tool use, deformable objects, and articulated objects, we demonstrate KALIEcan robustly solve the task and consistently outperform baselines. We hope KALIE will inspire future research towards adapting vision-language models for open-world robotic control.

Limitations and future work. The current affordance representation in KALIE is still limited to single-arm table-top manipulation. To apply KALIE to more complicated scenarios, such as dynamic manipulation and whole-body control, we would need to extend the design of the affordance representation. In addition, there is still a large discrepancy between the open-source VLM fine-tuned by KALIE and the state-of-the-art VLMs without accessible fine-tuning APIs. In the future, we hope to apply KALIE to more powerful VLMs to further improve the performance.

References

[1] S. Levine, P. Pastor, A. Krizhevsky, and D. Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research*, 37:421 – 436, 2016. URL https://api.semanticscholar.org/CorpusID:13072941.

[2] L. Pinto and A. K. Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3406–3413, 2015. URL https://api.semanticscholar.org/CorpusID:3177253.

[3] A. Brohan et al. Rt-1: Robotics transformer for real-world control at scale, 2023.

[4] OpenAI. Gpt-4 technical report, 2024.

[5] G. Team. Gemini: A family of highly capable multimodal models, 2024.

[6] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song, J. Xu, B. Xu, J. Li, Y. Dong, M. Ding, and J. Tang. Cogvlm: Visual expert for pretrained language models, 2024.

[7] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023.

[8] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng. Code as policies: Language model programs for embodied control. In *IEEE International Conference on Robotics and Automation*, pages 9493–9500. IEEE, 2023.

[9] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023.

[10] K. Fang, F. Liu, P. Abbeel, and S. Levine. MOKA: Open-Vocabulary Robotic Manipulation through Mark-Based Visual Prompting. In *Conference on Robot Learning (CoRL)*, 2024.

[11] D. Shah, B. Osinski, B. Ichter, and S. Levine. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on Robot Learning*, 2022. URL https://api.semanticscholar.org/CorpusID:250426345.

[12] A. Brohan et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023.

[13] Octo Model Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, C. Xu, J. Luo, T. Kreiman, Y. Tan, D. Sadigh, C. Finn, and S. Levine. Octo: An open-source generalist robot policy. https://octo-models.github.io, 2023.

[14] E. Collaboration. Open x-embodiment: Robotic learning datasets and rt-x models, 2024.

[15] A. Khazatsky et al. Droid: A large-scale in-the-wild robot manipulation dataset, 2024.

[16] H. Walke, K. Black, A. Lee, M. J. Kim, M. Du, C. Zheng, T. Zhao, P. Hansen-Estruch, Q. Vuong, A. He, V. Myers, K. Fang, C. Finn, and S. Levine. Bridgedata v2: A dataset for robot learning at scale, 2024.

[17] L. Manuelli, W. Gao, P. R. Florence, and R. Tedrake. kpam: Keypoint affordances for category-level robotic manipulation. In *International Symposium of Robotics Research*, 2019. URL https://api.semanticscholar.org/CorpusID:80628296.

[18] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[19] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

[20] T. Yu, T. Xiao, A. Stone, J. Tompson, A. Brohan, S. Wang, J. Singh, C. Tan, M. Dee, J. Peralta, B. Ichter, K. Hausman, and F. Xia. Scaling robot learning with semantically imagined experience. *ArXiv*, abs/2302.11550, 2023. URL https://api.semanticscholar.org/CorpusID:257079001.

[21] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. B. Girshick. Segment anything. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3992–4003, 2023. URL https://api.semanticscholar.org/CorpusID:257952310.

[22] X. Soria, Y. Li, M. Rouhani, and A. D. Sappa. Tiny and efficient model for the edge detection generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1364–1373, October 2023.

[23] W. Chen, O. Mees, A. Kumar, and S. Levine. Vision-language models provide promptable representations for reinforcement learning, 2024.

[24] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2023.

[25] W. Hong, W. Wang, Q. Lv, J. Xu, W. Yu, J. Ji, Y. Wang, Z. Wang, Y. Zhang, J. Li, B. Xu, Y. Dong, M. Ding, and J. Tang. Cogagent: A visual language model for gui agents, 2023.

[26] Y. Zhai, H. Bai, Z. Lin, J. Pan, S. Tong, Y. Zhou, A. Suhr, S. Xie, Y. LeCun, Y. Ma, and S. Levine. Fine-tuning large vision-language models as decision-making agents via reinforcement learning, 2024.

[27] B. Chen, C. Shu, E. Shareghi, N. Collier, K. Narasimhan, and S. Yao. Fireact: Toward language agent fine-tuning, 2023.

[28] A. Zeng, M. Liu, R. Lu, B. Wang, X. Liu, Y. Dong, and J. Tang. Agenttuning: Enabling generalized agent abilities for llms, 2023.

[29] J. Yang, H. Zhang, F. Li, X. Zou, C. Li, and J. Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023.

[30] P. Wang, S. Wang, J. Lin, S. Bai, X. Zhou, J. Zhou, X. Wang, and C. Zhou. One-peace: Exploring one general representation model toward unlimited modalities, 2023.

[31] K. Chen, Z. Zhang, W. Zeng, R. Zhang, F. Zhu, and R. Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic, 2023.

[32] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. H. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, and P. R. Florence. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning*, 2023. URL https://api.semanticscholar.org/CorpusID:257364842.

[33] M. Shridhar, L. Manuelli, and D. Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on Robot Learning*, pages 894–906. PMLR, 2022.

[34] Y. Hu, F. Lin, T. Zhang, L. Yi, and Y. Gao. Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning. *arXiv preprint arXiv:2311.17842*, 2023.

[35] I. Kostrikov, D. Yarats, and R. Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. *ArXiv*, abs/2004.13649, 2020. URL https://api.semanticscholar.org/CorpusID:216562627.

[36] M. Laskin, K. Lee, A. Stooke, L. Pinto, P. Abbeel, and A. Srinivas. Reinforcement learning with augmented data. *Advances in neural information processing systems*, 33:19884–19895, 2020.

[37] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017.

[38] B. Mehta, M. Diaz, F. Golemo, C. J. Pal, and L. Paull. Active domain randomization. In *Conference on Robot Learning*, pages 1162–1176. PMLR, 2020.

[39] I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, B. McGrew, A. Petron, A. Paino, M. Plappert, G. Powell, R. Ribas, et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

[40] H. Qi, B. Yi, S. Suresh, M. Lambeta, Y. Ma, R. Calandra, and J. Malik. General in-hand object rotation with vision and touch. In *Conference on Robot Learning*, pages 2549–2564. PMLR, 2023.

[41] A. Kar, A. Prakash, M.-Y. Liu, E. Cameracci, J. Yuan, M. Rusiniak, D. Acuna, A. Torralba, and S. Fidler. Meta-sim: Learning to generate synthetic datasets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4551–4560, 2019.

[42] Y. Zhang, H. Ling, J. Gao, K. Yin, J.-F. Lafleche, A. Barriuso, A. Torralba, and S. Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10145–10155, 2021.

[43] Y. Wang, Z. Xian, F. Chen, T.-H. Wang, Y. Wang, K. Fragkiadaki, Z. Erickson, D. Held, and C. Gan. Robogen: Towards unleashing infinite data for automated robot learning via generative simulation. *arXiv preprint arXiv:2311.01455*, 2023.

[44] M. Yang, Y. Du, K. Ghasemipour, J. Tompson, D. Schuurmans, and P. Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023.

[45] J. Devaranjan, A. Kar, and S. Fidler. Meta-sim2: Unsupervised learning of scene structure for synthetic data generation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 715–733. Springer, 2020.

[46] E. Wood, T. Baltrušaitis, C. Hewitt, S. Dziadzio, T. J. Cashman, and J. Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3681–3691, 2021.

[47] S. W. Kim, B. Brown, K. Yin, K. Kreis, K. Schwarz, D. Li, R. Rombach, A. Torralba, and S. Fidler. Neuralfield-ldm: Scene generation with hierarchical latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8496–8506, 2023.

[48] D. Li, H. Ling, S. W. Kim, K. Kreis, S. Fidler, and A. Torralba. Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21330–21340, 2022.