2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Open-World Object Manipulation with Vision-Language-Action Models via Synthetic Multi-Modal Data

Yefei Chen[1], Junjie Wen[1], Jinming Li[2], Zhongyi Zhou[1],
Yaxin Peng[2], Chaomin Shen[1,†], Yi Xu[4], Yichen Zhu[3,†]

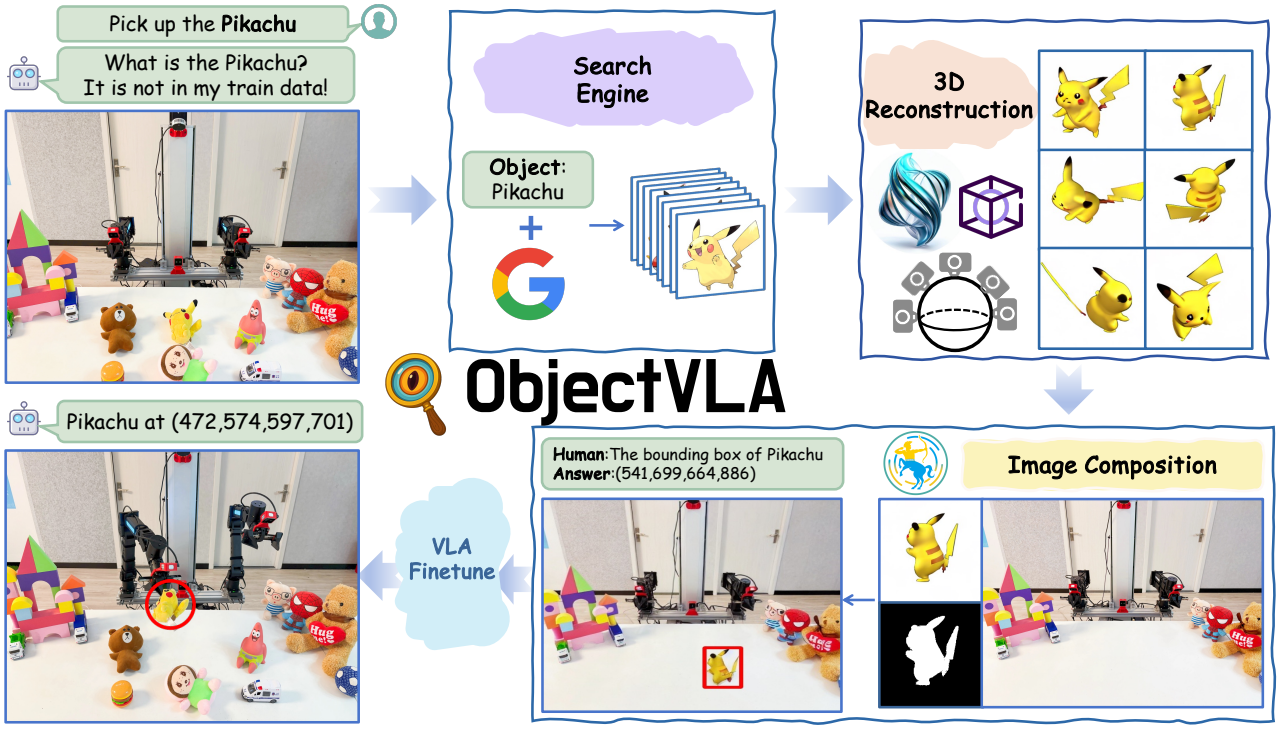

Fig. 1: **Overview of the ObjectVLA Framework.** When encountering an unseen object (e.g., Pikachu), ObjectVLA searches the web for images of the object, generates multi-view 3D representations, and composes them into contextual scenes. After VLA finetuning, the robot can accurately localize and interact with the object.

***Abstract*— Imitation learning has proven to be highly effective in teaching robots dexterous manipulation skills. However, it typically relies on large amounts of robot data, which limits its scalability and applicability in dynamic, real-world environments. One key challenge in this context is object generalization—where a robot trained to perform a task with one object, such as "hand over the apple." struggles to transfer its skills to a semantically similar but visually different object, such as "hand over the peach." This gap in generalization to new objects beyond those in the same category has yet to be adequately addressed in previous work on end-to-end visuomotor policy learning. In this paper, we present a simple yet effective approach for achieving object generalization through Vision-Language-Action (VLA) models, referred to as ObjectVLA. We design a lightweight image-text-data-synthesis pipeline, Search2Scene, which enables robots to generalize learned skills to novel objects without requiring explicit human demonstrations for each new target object. By leveraging vision-language pair data, our method provides a lightweight and scalable way to inject knowledge about the target object, establishing an implicit link between the object and the desired action. We evaluate ObjectVLA on a real robotic platform, demonstrating its ability to generalize across 100 novel objects with a 64% success rate in selecting objects not seen during training. These results highlight the effectiveness of our approach in enabling object-level generalization and reducing the need for extensive human demonstrations, paving the way for more flexible and scalable robotic learning systems.**

[1]School of Computer Science, East China Normal University, [2]Department of Mathematics, School of Science, Shanghai University, [3]University of Toronto [4] Midea Group
[†]Corresponding authors
This work was done while Yefei Chen, Junjie Wen, Jinming Li, Zhongyi Zhou and Yichen Zhu were at Midea Group.

I. INTRODUCTION

Vision-language-action (VLA) models have emerged as a transformative paradigm for teaching robots dexterous skills, enabling them to replicate human behavior and master complex tasks [1]–[6]. However, a critical limitation persists: these models rely heavily on human demonstration data, which constrains their scalability and practicality in dynamic

real-world environments [7]–[9]. For instance, a robot trained to execute "hand over the apple." often fails to generalize to analogous tasks like "hand over the peach." despite conceptual similarity. This underscores the unresolved challenge of **object generalization** — adapting learned skills to novel, unseen objects — particularly when such objects lie **beyond the category of the teleoperated training data**. We name these objects as out-of-distribution (OOD) objects.

The core limitation stems from imitation learning's tendency to learn fixed mappings from instruction and visual input to action. When encountering objects absent from teleoperation data, the model lacks mechanisms to associate the object's name, visual features, and learned actions. To address this, we propose a framework that bridges visual-language semantics and robotic actions through automatically synthesized image-text data and localization-aware reasoning.

To support zero-shot generalization, we introduce Search2Scene, a dedicated pipeline for synthesizing high-quality image–text data. Search2Scene first parses user input to identify a target object and searches for high-quality images of it. The pipeline then processes these images using 3D reconstruction to create multi-view representations. These representations are subsequently rendered into background scenes to synthesize composite images. Finally, the resulting composite images, along with associated localization annotations, are paired with text descriptions to form a structured image–text dataset. This dataset is co-finetuned with teleoperated robot interaction data, while the robot data itself is enriched with localization-guided reasoning. By using this pipeline to produce data localization as a bridging representation, we create a unified pathway between visual-language inputs and robotic actions. This enables zero-shot object generalization: the model can recognize and manipulate novel objects—even those absent from robot training data—without task-specific retraining.

We designed rigorous real-robot experiments to validate the effectiveness of image-text data augmented with localization metadata (e.g., bounding boxes). With six objects positioned on either side of a table, the robot followed "move to the object" commands, scoring 100% on in-domain items and 64% on 100 OOD objects. The versatility of our approach is further demonstrated across diverse scenarios, including bin-picking task and other tasks requiring composite skills like pushing and rotating. These experiments highlight the importance of synthesizing accurate and realistic image-text data, which enhances generalization and reduces reliance on real-world robotic data.

Our primary contribution is a unified pipeline that automatically synthesizes image-text datasets and integrates them with robot interaction data, enabling end-to-end object generalization. Designed to be modular and scalable, the framework can be easily extended to handle different objects within the same task, without requiring additional robot demonstrations. Despite some of the existing works, such as RT-2 [2] and ECoT [5] giving a glimpse of how co-finetuning can achieve simple object generalization, they neither elucidate the underlying mechanism of achieving such generalization nor address the boundary of their methodologies. In contrast, our approach — though simple and straightforward — demonstrates that training VLA models with a hybrid dataset of robot interaction data and image-text data significantly enhances generalization. This level of generalization goes significantly beyond previously demonstrated end-to-end approaches. Crucially, our framework enables practical deployment: simply searching for an object by name to synthesize the corresponding dataset and quickly align it with existing robot skills—achieving zero-shot generalization.

II. Related Work

Vision-language-action models for robot control. Recent research has focused on developing generalist robot policies trained on increasingly expansive robot learning datasets [10]–[14]. Vision-language-action models (VLAs) represent a promising approach for training such generalist policies [3], [4], [6], [7], [15]–[20]. VLAs adapt vision-language models (VLMs) [21]–[30], pre-trained on internet-scale image and text data, for robotic control. This approach offers several advantages: leveraging large vision-language model backbones, with billions of parameters, provides the necessary capacity for fitting extensive robot datasets. Furthermore, reusing weights pre-trained on internet-scale data enhances the ability of VLAs to interpret diverse language commands and generalize to novel objects and environments. However, current VLA models struggle to recognize open-world objects when these objects absent from the robot interaction data [7], [8]. This is mainly due to VLMs essentially "overwrites" its previously acquired knowledge of open-world objects with robot-specific information.

Generalization in robot learning. In the realm of robot learning, generalization, particularly object generalization, remains a core challenge and active area of research. Many works leverage techniques such as domain randomization [31], meta-learning [32], [33], retrieval-augmented generation [34], extra modality [35], [36], and data augmentation to improve a robot's ability to recognize and interact with novel objects unseen during training. For instance, domain randomization methods [31], [37] randomize visual and physical parameters during simulation training to force the agent to learn features invariant to these irrelevant details, leading to better real-world generalization. Furthermore, meta-learning approaches [38] aim to train models that can rapidly adapt to new objects with limited data, directly addressing the object generalization problem. Finally, data augmentation methods [39], [40], enhance the diversity of the training data, exposing the model to a wider range of object appearances and orientations, thereby promoting robustness and generalization to novel objects. There is also a field of work using large language models or vision-language models to do open-vocabulary manipulation [41]–[44], combined with motion planning and robot learning methods. However, these approaches involve separate modules that are trained independently for different components. To the best of our knowledge, this work represents the first exploration

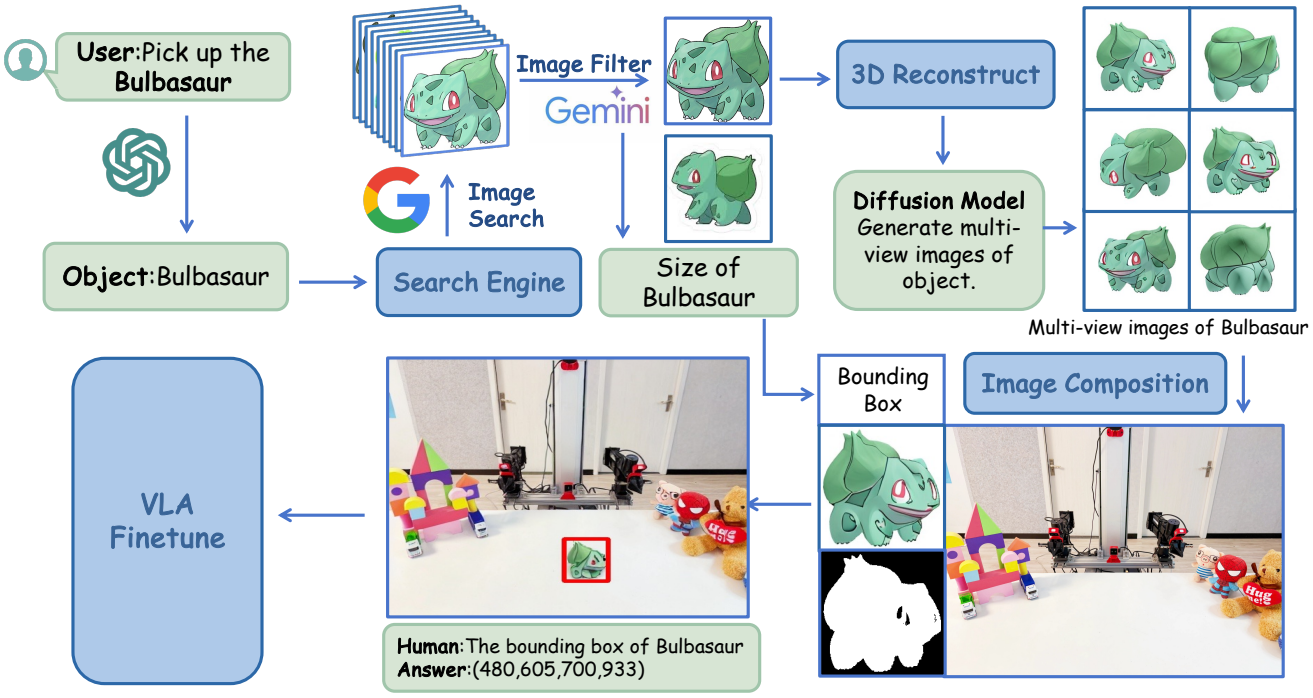

Fig. 2: **Overview of the Search2Scene** Given an object name (e.g., Bulbasaur), the system searches for web images, uses Gemini to filter them and to estimate object size, and generates multi-view 3D images via a diffusion model. These are composed into scenes with auto-generated bounding boxes for VLA finetuning.

of object generalization beyond specific categories within visuomotor policy learning, and further proposes a plug-and-play module to address this generalization challenge.

III. Methodology

A. Notation and Motivation

Given a set of expert demonstrations that contain complex robot skill trajectories, we want to learn a visuomotor policy $\pi : \{\mathcal{O}_r, \mathcal{I}_r\} \mapsto \mathcal{A}$ that maps the visual observations $o_r \in \mathcal{O}_r$ and the language instruction $i_r \in \mathcal{I}_r$ to actions $a \in \mathcal{A}$. The action changes accordingly when the language instruction and visual input change. The r denotes the data in the human demonstration data. Typically, for each language instruction it contains robot skill such as "push" or "pick up" and the target object, which is denoted as $\{obj_r, skill_r\} \in i_r$. We then formally define the image-text data, where $\varphi : \{\mathcal{O}_v, \mathcal{I}_v\} \mapsto \mathcal{L}_v$, where we input the image $o_v \in \mathcal{O}_v$ and give a language instruction $i_v \in \mathcal{I}_v$, the model is output with the corresponding answer $l_r \in \mathcal{L}_v$. The notation v denotes image-text data. In this work, we explore the generalization of objects, focusing on those that are not part of the robot interaction data but are present in image-text data.

B. Search2Scene

In this section, we present the overall framework of Search2Scene, and detail the motivations and specifics of three key modules: 3D Reconstruction, Image Composition, and Data Construction.

ObjectVLA is an end-to-end VLA model augmented with Search2Scene to enhance its generalization ability. While existing methods rely heavily on human-collected demonstration data for training, which limits their scalability in dynamic and unstructured environments. In this work, Search2Scene processes user queries to search relevant images and composes multi-view sequences, enabling the VLA model to generalize to novel objects in zero-shot settings. An overview of the Search2Scene workflow is illustrated in Figure 2, and the following sections detail each component of the system.

1) Query analysis: GPT-4o [45] analyzes the user instruction to extract the object to be searched.

2) Search Engine: Given a target object name, we first query the Google Search to obtain 10 candidate images. Next, Gemini-2.0-flash-exp [46] executes a filtering step, selecting images based on several criteria: semantic alignment with the object name, color matching, the presence of a single object instance, and a clean or transparent background. This process yields a refined set of 2–3 high-quality images for 3D Reconstruction.

3) 3D Reconstruction: Directly using Google search results for image composition presents several challenges. First, high-quality images that match the query are limited in number. As you search for more images, you get a lot more bad ones, which hurts the quality of the synthesized images. Second, the searched object views are typically constrained to a single perspective, lacking multi-view visual diversity and thus limiting the model's holistic understanding of the object. Third, the searched images exhibit significant variation in resolution and background complexity. These differences make it even harder to composite images.

To address the challenges posed by noisy image data, we first apply an image filtering step to remove low-quality samples. However, this results in an insufficient number of images for training. Therefore, we use a multi-view data generation method based on 3D reconstruction. First, using the images filtered by Gemini, we perform 3D reconstruction with the One-2-3-45 [47] framework, obtaining 32 novel view images from different perspectives. This effectively expanded the original dataset and enriched the object's viewpoint information. It's important to note that due to the limited quality of the generated 3D models, we don't use the reconstructed models to obtain multi-view images. Instead, we directly select the multi-view images generated by the model during the 3D reconstruction process as subsequent data. This underscores that our approach is not only time-efficient but also yields higher quality results.

4) Image Composition: To enhance the model's generalization ability in unfamiliar environments, we aim to synthesize composite images containing the target object integrated into novel scenes. Therefore, we employ ControlCom [48] as our image composition method. ControlCom is a controllable image composition method based on diffusion models, capable of unifying four tasks—image blending, image harmonization, view synthesis, and generative composition—within a single model, while also improving foreground fidelity.

First, we extract the target object from multi-view images and generate corresponding mask images. To realistically reflect the object's size in the composite image, we use Gemini to obtain the object's typical size and represent it with a bounding box. Considering that excessively small bounding boxes lead to poor synthesis results, we set a lower limit for the bounding box size. We input the extracted foreground image, mask image, background image, and bounding box into the ControlCom model to synthesize a new composite image. The mask image guides the fusion area between the foreground and background, ensuring a natural visual transition of the synthesized object within the background environment. After composition, the foreground object is accurately placed onto the background image according to the bounding box position.

5) Data Construction:

Image-text data construction. To enhance the model's ability to generalize to novel objects, we construct a diverse image-text dataset. For the visual component, we use the composite images. In the experiment that demonstrates the necessity of image-text data augmented with localization metadata, we collected 20 distinct camera-captured views for each of 100 objects. For the textual component, we employ a fixed template, "Detecting the bounding box of object.", as the question, and the bounding box as the answer.

Reasoning data construction. We utilize localization metadata to bridge the gap between image-text data and robot data, as previously mentioned. To establish this implicit link between image-text and action, we incorporate localization metadata into the robot data. This section details how we construct reasoning with localization for robot data.

For each task, we first identify target objects based on the language instructions. We then employ DinoX [49], a cutting-edge open-vocabulary object detector, to annotate the bounding boxes of these objects. DinoX can generate a bounding box given an object's name. To ensure accuracy, we manually verify and correct any erroneous bounding boxes produced by DinoX. Since our workspace has two external camera views, which can result in different bounding boxes for the same object, we annotate only one (right camera in our experiments). Following Qwen2-VL [25], we use a fixed template, "<|object_ref_start|>{object}<|object_ref_end|><|box_start|>(x_1, y_1),(x_2, y_2)<|box_end|>.", to represent the localization reasoning. This reasoning is generated before each action and injected into the policy model through a learnable module. For a detailed explanation of this injection module's architecture, we refer readers to DiffusionVLA [9], the base model used in our experiments.

IV. Experiments

In this section, we examine the effectiveness of ObjectVLA for object generalization in embodied control. In section IV-A, we verify the effectiveness of our method in object generalization. In section IV-B and IV-C, we illustrate how our model transfers skills to objects not present in robot data but included in the vision-language corpus.

Real robot setup. Our experiments are conducted on a Franka robot [50] equipped with a 7-degree-of-freedom arm and a gripper. We use two external ZED cameras and a wrist Realsense 435i camera to obtain real-world visual information. Our real-world robot setups are illustrated in Figure 3.

A. Validating Object Generalization

In this section, we conduct rigorous experiments to verify the object generalization capability of our method. We begin by describing the experimental setup and evaluation criteria. Next, we evaluate ObjectVLA on both in-distribution and out-of-distribution objects. Finally, we explore several interesting observations related to object generalization.

1) Experimental Setup: To verify the object generalization capability, we begin with a simple yet effective task, "Move to the object.". In this task, we position objects on both sides of the robot, ensuring that each side has at least three objects on the table. The model is required to move toward the target object based on the given instruction. These objects are randomly chosen from a diverse set. For in-distribution (ID) evaluation, objects are only selected from the robot's training data. And, for out-of-distribution evaluation, objects are randomly selected from either the robot's training data or the vision-language data. A complete list of objects from both datasets is provided in the Appendix.

Evaluation criterion. We evaluate each object over 4 trials, with the target area's side switching every two trials. We consider the model to have successfully recognized a novel object if and only if it moved toward the target object in all four trials. This criterion ensures that the model cannot achieve success simply by chance.

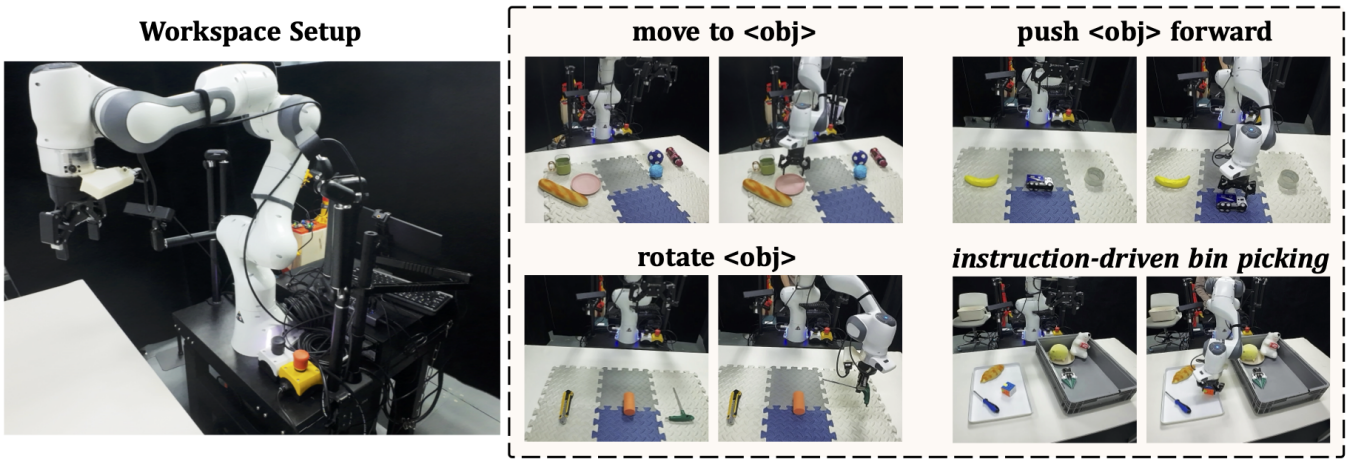

Fig. 3: **Robot setup and examples for real-world manipulation tasks.** We evaluate ObjectVLA with 4 skills on a Franka robot arm equipped with two external Zed cameras and a Realsense 435i wrist camera.

Experimental results. Figure 5 presents the real-world experimental results for the "Move" task. Our ObjectVLA achieves a 100% success rate in ID evaluation, surpassing the strong π_0 baseline by ten percentage points (90%). Under the stress-test regime with out-of-distribution (OOD) objects, it preserves a 64% success rate—an gain of 20 percentage points over π_0's 44%. These gains confirm the effectiveness of co-training robot data with localization metadata.

Ablation study. To further understand our method's effectiveness, we conducted an ablation study. We found that object generalization relies heavily on two key factors: first, explicitly linking vision and language to action through bounding boxes. This provides a direct connection between the visual object, its linguistic description, and the required manipulation. Second, designing a reasoning process in the robot data that mirrors the structure of vision-language pair data. This allows the model to leverage the rich information encoded in pre-trained vision-language models.

To analyze the impact of these factors, we removed the reasoning module for robot data and eliminated bounding boxes for vision-language data. The VLA model is then co-finetuned with vision-language data and evaluated using the same criteria and test settings as our full method.

As illustrated in Figure 5, the model without bounding boxes achieves only a 19% success rate in OOD evaluation, representing a significant performance decline compared to our method, despite achieving a 100% success rate in the ID test. This suggests that without explicit grounding and a structured reasoning process, the model struggles to differentiate objects in vision-language data, leading to confusion about object-instruction correspondence and appropriate action selection.

2) More Observations:

Can VLA recognize unseen objects if only trained with teleoperated data? To further assess the importance of vision-language data, we evaluated a VLA model trained exclusively on robot data, without any vision-language co-finetuning. As shown in Figure 5, this model (DiVLA) achieved 8% accuracy, which is almost equivalent to random guessing. This stark outcome highlights the critical role of vision-language data in multimodal understanding.

TABLE I: **Experimental Results for rotate and push skills.** Our proposed ObjectVLA achieves high performance on both 5 in-distribution objects and 20 out-of-distribution objects. Each object is evaluated with three trials. We report the number of success trials.

Task	Method	ID	OOD
Rotate	π_0	10/15	22/60
	ObjectVLA	**13/15**	**39/60**
Push	π_0	12/15	32/60
	ObjectVLA	**12/15**	**52/60**

While the VLA model's backbone is pre-trained on internet-scale vision-language data, focusing solely on robot data during training leads to catastrophic forgetting. The model essentially "overwrites" its previously acquired knowledge of visual concepts with robot-specific information, hindering its ability to comprehend multimodal scenes. Consequently, even objects encountered during pre-training, such as Pikachu, remain unrecognizable to the VLA model without vision-language co-finetuning.

B. Combining with More Skills

While the previous section employed a simple "move to" demonstration to validate the fundamental approach of our method, this section expands the evaluation to encompass more complex skills, specifically "push" and "rotate." This assessment aims to demonstrate the generalizability of our method and its applicability beyond the "move to" task.

Experimental setup. In this experimental setup, we placed three objects in front of the robot: one on the center, one on the right, and one on the left. The robot is instructed to either "rotate the object counterclockwise" or "push the object forward," as illustrated in Figure 3. Following previous setup,

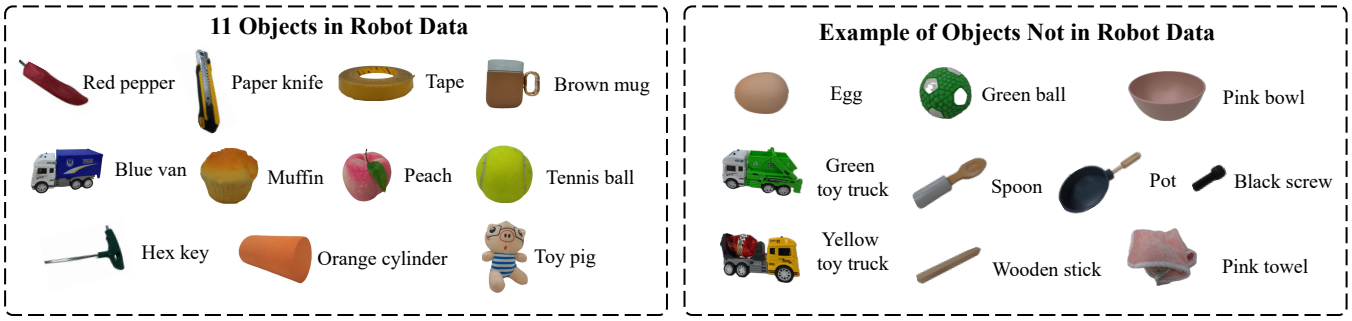

Fig. 4: **Example Objects Used in Experiments.** *Left:* Objects present in the robot training data. *Right:* Examples of novel objects, not present in the robot data, but included in the image-text co-training dataset.

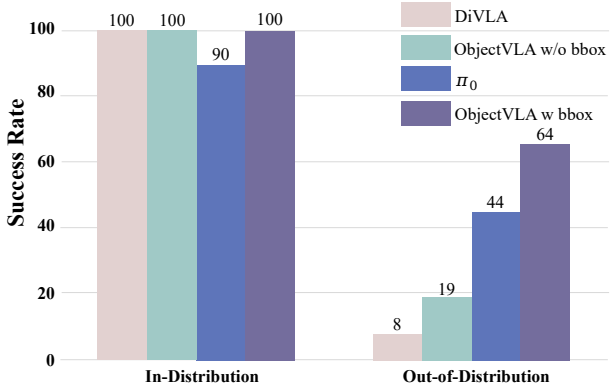

Fig. 5: **Validation experiments on object generalization.** Our method achieved the best performance in both the in-distribution test setup and under visual changes. Each object is evaluated across 4 trials. We report the number of objects that were correctly identified in all four trials.

we evaluate the model's performance for both in-distribution (ID) and out-of-distribution (OOD) objects. Recognizing that some objects are inherently unsuitable for rotation or pushing actions (e.g., dishes), we conducted experiments on a curated set of 5 ID objects and 20 OOD objects. For each object in a skill, 40 demonstrations were collected, resulting in a total of 400 demonstrations.

Implementation details. We train one model for each skill to ensure that the model focuses more on understanding the objects rather than multi-task learning. We use the same image-textual data of "move" task. Following established protocols from our prior work, this image-text dataset trained concurrently with the demonstration data for comprehensive evaluation. Each object was tested with 3 trials. In total, 150 trials were conducted.

Results. As shown in Table I, our method achieved high success rates on the robot interaction objects for both rotate and push skill. More importantly, when evaluated on out-of-distribution objects, it exhibits a markedly stronger performance than the π_0 baseline. Analysis of the failed "rotate" trials revealed that the primary cause is the model's inability to grasp the target object securely. When evaluating performance on out-of-distribution (OOD) objects,

TABLE II: **Experimental results for bin picking.** Our proposed ObjectVLA achieves high performance on both 11 in-distribution objects and 50 out-of-distribution objects, with each object evaluated across 3 trials. We report the number of successful trials over total trials.

Method	ID	OOD
OpenVLA	14/33	17/150
π_0	19/33	62/150
ObjectVLA	**21/33**	**87/150**

we observed a decrease in task completion rates compared to in-distribution objects, as expected. However, the model still successfully completed nearly two-thirds of the trials. Notably, in most failure cases, the model did not incorrectly identify the target object but rather failed to execute the skill completely. This was particularly evident in the "rotate" trials, where successful execution hinges on a secure grasp, a challenging requirement for unseen objects. Nevertheless, these experiments strongly support the claim that ObjectVLA can transfer learned skills, beyond basic pick and place, to novel objects within the framework we have developed. The results underscore the potential of ObjectVLA for generalized robotic manipulation, capable of adapting to new objects and tasks beyond its initial training.

C. Instruction-Driven Bin Picking

To further evaluate ObjectVLA, we conducted experiments in a more practical scenario: end-to-end instruction-driven bin-picking. Unlike prior works (e.g., GR-2 [51] and DiVLA [9]) that execute bin-picking tasks without specific semantic instructions—typically limited to generic actions like transferring all objects from one container to another—we focus on a significantly more challenging setting [9], [51]. In our experiments, the robot is required to identify and retrieve a specific target object based on natural language instructions (e.g., "Pick the hexagonal bolt from the bin"). This scenario elevates the complexity of conventional bin-picking tasks by integrating cross-modal understanding (vision-to-language alignment) and fine-grained object discrimination that have multiple objects in the scene. Notably, the objects are randomly placed on the panel, which is a large area. Not

only does the model need to figure out the object's position, but also needs to be aware of its pose.

Implementation details. We collected new data within this environment. For robot data, we collected 600 pick-and-place trajectories using the same "seen" objects as in previous experiments. For image-text data, we used half the number of objects from previous experiments, capturing 20 images of each. We compared our method against OpenVLA, a state-of-the-art VLA model, reporting success rates for both in-distribution and out-of-distribution objects. Evaluation consisted of three trials per object, totaling 183 trials per method. In each trial, at least two objects were randomly placed on the plate, and the model was instructed to pick and place a specific object according to the given instruction.

Results. Table II presents our results. Bin picking, requiring object retrieval from random positions and poses, poses a significant challenge even for in-distribution objects. OpenVLA achieves a success rate of only 42.4% for in-distribution objects, significantly less than half. The stronger π_0 raises this to 57.6%. Surprisingly, OpenVLA still completes about 11% of trials with out-of-distribution objects, while π_0 manages 41.3%, likely because some test items share attributes with training objects (e.g., bread resembling a muffin, a green mug differing only in color from a brown training mug). In contrast, our method completes 63.6% of in-distribution trials and 58.0% of out-of-distribution trials—an absolute improvement of 21.3% over π_0 and 46.7% over OpenVLA on OOD objects. These results further emphasize the necessity of co-training with both robot interaction and image-text data for effective object generalization.

V. Conclusion

In this work, we present ObjectVLA, a unified Vision-Language-Action framework that addresses object generalization in robotic manipulation. By co-finetuning a policy on this synthetic data alongside demonstrations, our method establishes Search2Scene that bridges semantic understanding and physical action execution. This enables zero-shot generalization to over 100 novel objects with a 64% success rate, even when objects differ in category, appearance, or fine-grained attributes (e.g., color, shape). Our framework demonstrates that lightweight co-training with image-textual priors and localization-aware reasoning can unlock robust cross-modal alignment. Key to our success is the ability to adapt rapidly to real-world scenarios: by simply searching the web for images of a target object, generating multi-view 3D representations, and compositing them into contextual scenes, we can accumulate large volumes of realistic image–text data for skill transfer. After quick continual finetuning, robots generalize to unseen objects without costly human demonstrations. Our results highlight a path toward scalable robotic learning systems that reduce dependence on large-scale teleoperation data while maintaining high performance.

VI. Limitation

There are still a number of limitations in this work. One key issue is that image-text data sourced from the internet may differ significantly from the appearance of target objects in real-world robot manipulation tasks. Such discrepancies could hinder the effectiveness of skill transfer, especially when the visual domain gap is large. In particular, transferring skills to objects that lack distinctive visual features tends to be more challenging and less reliable. Our primary focus here is to introduce a novel pipeline that enables deep learning models to transfer skills to new objects without explicit demonstrations. Determining the limits of this transferability, particularly concerning the permissible visual gap between training and target objects, remains an open question for future investigation. Currently, our method struggles to generalize to novel backgrounds and lighting conditions. We believe the visual gap between our collected image-text data and the robot's operational environment contributes to this challenge. Bridging this gap to improve generalization is a key focus for future development.

Acknowledgment

This work is supported by the Sci-Tech Innovation Initiative by the Science and Technology Commission of Shanghai Municipality (24ZR1419000), the National Science Foundation of China (12471501)and ECNU Multifunctional Platform for Innovation (001).

References

[1] 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] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv preprint arXiv:2307.15818*, 2023.

[3] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, L. X. Shi, J. Tanner, Q. Vuong, A. Walling, H. Wang, and U. Zhilinsky, "π_0: A vision-language-action flow model for general robot control," 2024. [Online]. Available: https://arxiv.org/abs/2410.24164

[4] K. Pertsch, K. Stachowicz, B. Ichter, D. Driess, S. Nair, Q. Vuong, O. Mees, C. Finn, and S. Levine, "Fast: Efficient action tokenization for vision-language-action models," *arXiv preprint arXiv:2501.09747*, 2025.

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

[6] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai *et al.*, "pi0.5: a vision-language-action model with open-world generalization," *arXiv preprint arXiv:2504.16054*, 2025.

[7] M. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[8] J. Wen, Y. Zhu, J. Li, M. Zhu, K. Wu, Z. Xu, R. Cheng, C. Shen, Y. Peng, F. Feng *et al.*, "Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation," *arXiv preprint arXiv:2409.12514*, 2024.

[9] J. Wen, M. Zhu, Y. Zhu, Z. Tang, J. Li, Z. Zhou, C. Li, X. Liu, Y. Peng, C. Shen *et al.*, "Diffusion-vla: Scaling robot foundation models via unified diffusion and autoregression," *arXiv preprint arXiv:2412.03293*, 2024.

[10] A. O'Neill, A. Rehman, A. Gupta, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models," *arXiv preprint arXiv:2310.08864*, 2023.

[11] 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.

[12] H.-S. Fang, H. Fang, Z. Tang, J. Liu, J. Wang, H. Zhu, and C. Lu, "Rh20t: A robotic dataset for learning diverse skills in one-shot," *arXiv preprint arXiv:2307.00595*, 2023.

[13] S. Dasari, O. Mees, S. Zhao, M. K. Srirama, and S. Levine, "The ingredients for robotic diffusion transformers," *arXiv preprint arXiv:2410.10088*, 2024.

[14] F. Lin, Y. Hu, P. Sheng, C. Wen, J. You, and Y. Gao, "Data scaling laws in imitation learning for robotic manipulation," 2024. [Online]. Available: https://arxiv.org/abs/2410.18647

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

[16] J. Wen, Y. Zhu, J. Li, Z. Tang, C. Shen, and F. Feng, "Dexvla: Vision-language model with plug-in diffusion expert for general robot control," *arXiv preprint arXiv:2502.05855*, 2025.

[17] D. Niu, Y. Sharma, G. Biamby, J. Quenum, Y. Bai, B. Shi, T. Darrell, and R. Herzig, "Llarva: Vision-action instruction tuning enhances robot learning," *arXiv preprint arXiv:2406.11815*, 2024.

[18] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *arXiv preprint arXiv:2303.04137*, 2023.

[19] Z. Zhou, Y. Zhu, M. Zhu, J. Wen, N. Liu, Z. Xu, W. Meng, R. Cheng, Y. Peng, C. Shen *et al.*, "Chatvla: Unified multimodal understanding and robot control with vision-language-action model," *arXiv preprint arXiv:2502.14420*, 2025.

[20] Z. Zhou, Y. Zhu, J. Wen, C. Shen, and Y. Xu, "Vision-language-action model with open-world embodied reasoning from pretrained knowledge," *arXiv preprint arXiv:2505.21906*, 2025.

[21] M. Zhu, Y. Zhu, X. Liu, N. Liu, Z. Xu, C. Shen, Y. Peng, Z. Ou, F. Feng, and J. Tang, "Mipha: A comprehensive overhaul of multimodal assistant with small language models," *arXiv preprint arXiv:2403.06199*, 2024.

[22] H. Zhao, M. Zhang, W. Zhao, P. Ding, S. Huang, and D. Wang, "Cobra: Extending mamba to multi-modal large language model for efficient inference," *arXiv preprint arXiv:2403.14520*, 2024.

[23] Y. Zhu, M. Zhu, N. Liu, Z. Xu, and Y. Peng, "Llava-phi: Efficient multi-modal assistant with small language model," in *Proceedings of the 1st International Workshop on Efficient Multimedia Computing under Limited*, 2024, pp. 18–22.

[24] S. Karamcheti, S. Nair, A. Balakrishna, P. Liang, T. Kollar, and D. Sadigh, "Prismatic vlms: Investigating the design space of visually-conditioned language models," *arXiv preprint arXiv:2402.07865*, 2024.

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

[26] H. Lu, W. Liu, B. Zhang, B. Wang, K. Dong, B. Liu, J. Sun, T. Ren, Z. Li, H. Yang *et al.*, "Deepseek-vl: towards real-world vision-language understanding," *arXiv preprint arXiv:2403.05525*, 2024.

[27] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[28] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," *arXiv preprint arXiv:2310.03744*, 2023.

[29] M. Abdin, J. Aneja, H. Awadalla, A. Awadallah, A. A. Awan, N. Bach, A. Bahree, A. Bakhtiari, J. Bao, H. Behl *et al.*, "Phi-3 technical report: A highly capable language model locally on your phone," *arXiv preprint arXiv:2404.14219*, 2024.

[30] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu *et al.*, "Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 24 185–24 198.

[31] S. James, P. Wohlhart, M. Kalakrishnan, D. Kalashnikov, A. Irpan, J. Ibarz, S. Levine, R. Hadsell, and K. Bousmalis, "Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 627–12 637.

[32] G. Schoettler, A. Nair, J. A. Ojea, S. Levine, and E. Solowjow, "Meta-reinforcement learning for robotic industrial insertion tasks," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 9728–9735.

[33] R. Kaushik, T. Anne, and J.-B. Mouret, "Fast online adaptation in robotics through meta-learning embeddings of simulated priors," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 5269–5276.

[34] Y. Zhu, Z. Ou, X. Mou, and J. Tang, "Retrieval-augmented embodied agents," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 985–17 995.

[35] Y. Zhu, Z. Ou, F. Feng, and J. Tang, "Any2policy: Learning visuomotor policy with any-modality," *Advances in Neural Information Processing Systems*, vol. 37, pp. 133 518–133 540, 2025.

[36] W. Ye, F. Liu, Z. Ding, Y. Gao, O. Rybkin, and P. Abbeel, "Video2policy: Scaling up manipulation tasks in simulation through internet videos," *arXiv preprint arXiv:2502.09886*, 2025.

[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)*, 2017, pp. 23–30.

[38] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.

[39] M. Laskin, K. Lee, A. Stooke, L. Pinto, P. Abbeel, and A. Srinivas, "Reinforcement learning with augmented data," *Advances in neural information processing systems*, vol. 33, pp. 19 884–19 895, 2020.

[40] I. Kostrikov, D. Yarats, and R. Fergus, "Image augmentation is all you need: Regularizing deep reinforcement learning from pixels," *arXiv preprint arXiv:2004.13649*, 2020.

[41] A. Stone, T. Xiao, Y. Lu, K. Gopalakrishnan, K.-H. Lee, Q. Vuong, P. Wohlhart, S. Kirmani, B. Zitkovich, F. Xia *et al.*, "Open-world object manipulation using pre-trained vision-language models," *arXiv preprint arXiv:2303.00905*, 2023.

[42] Y. Zhu, A. Lim, P. Stone, and Y. Zhu, "Vision-based manipulation from single human video with open-world object graphs," *arXiv preprint arXiv:2405.20321*, 2024.

[43] 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 *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 9493–9500.

[44] 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," *arXiv preprint arXiv:2204.01691*, 2022.

[45] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[46] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen *et al.*, "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," *arXiv preprint arXiv:2507.06261*, 2025.

[47] M. Liu, C. Xu, H. Jin, L. Chen, M. Varma T, Z. Xu, and H. Su, "One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization," *Advances in Neural Information Processing Systems*, vol. 36, pp. 22 226–22 246, 2023.

[48] B. Zhang, Y. Duan, J. Lan, Y. Hong, H. Zhu, W. Wang, and L. Niu, "Controlcom: Controllable image composition using diffusion model," *arXiv preprint arXiv:2308.10040*, 2023.

[49] T. Ren, Y. Chen, Q. Jiang, Z. Zeng, Y. Xiong, W. Liu, Z. Ma, J. Shen, Y. Gao, X. Jiang *et al.*, "Dino-x: A unified vision model for open-world object detection and understanding," *arXiv preprint arXiv:2411.14347*, 2024.

[50] S. Haddadin, "The franka emika robot: A standard platform in robotics research [survey]," *IEEE Robotics Automation Magazine*, vol. 31, no. 4, pp. 136–148, 2024.

[51] C.-L. Cheang, G. Chen, Y. Jing, T. Kong, H. Li, Y. Li, Y. Liu, H. Wu, J. Xu, Y. Yang *et al.*, "Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation," *arXiv preprint arXiv:2410.06158*, 2024.