CONTROLMANIP: FEW-SHOT MANIPULATION FINE-TUNING VIA OBJECT-CENTRIC CONDITIONAL CONTROL

Puhao Li[1,2], **Yingying Wu**[1], **Wanlin Li**[2], **Yuzhe Huang**[2], **Zhiyuan Zhang**[2], **Yinghan Chen**[2,3], **Song-Chun Zhu**[1,2,3], **Tengyu Liu**[2†], **Siyuan Huang**[2†]
[†]Corresponding author [1]Department of Automation, Tsinghua University
[2]State Key Lab of General Artificial Intelligence, BIGAI [3]Institute for AI, Peking University

ABSTRACT

Learning real-world robotic manipulation is challenging, particularly when limited demonstrations are available. Existing methods for few-shot manipulation often rely on simulation-augmented data or pre-built modules like grasping and pose estimation, which struggle with sim-to-real gaps and lack versatility. While large-scale imitation pre-training shows promise, adapting these general-purpose policies to specific tasks in data-scarce settings remains unexplored. To achieve this, we propose ControlManip, a novel framework that bridges pre-trained manipulation policies with object-centric representations via a ControlNet-style architecture for efficient fine-tuning. Specifically, to introduce object-centric conditions without overwriting prior knowledge, ControlManip zero-initializes a set of projection layers, allowing them to gradually adapt the pre-trained manipulation policies. In real-world experiments across 6 diverse tasks, including pouring cubes and folding clothes, our method achieves a 73.3% success rate while requiring only 10-20 demonstrations — a significant improvement over traditional approaches that require more than 100 demonstrations to achieve comparable success. Comprehensive studies show that ControlManip improves the few-shot fine-tuning success rate by 252% over baselines and demonstrates robustness to object and background changes. By lowering the barriers to task development, ControlManip accelerates real-world robot adoption and lays the groundwork for unifying large-scale policy pre-training with object-centric representations.

1 INTRODUCTION

Robotic manipulation in the real world remains a fundamental challenge, particularly when learning novel skills from limited demonstrations. While recent advances in robotic manipulation (Fu et al., 2024; Li et al., 2024b; Liu et al., 2024; Zhu et al., 2023b; Chen et al., 2024; Wang et al., 2024b; Hsu et al., 2024; Chi et al., 2024; Ma et al., 2023; Jiang et al., 2024; Brohan et al., 2023; Yang et al., 2024; Li et al., 2025; Huang et al., 2024; Li et al., 2023; 2024c) have shown promise, current methods still demand extensive training data and struggle to efficiently adapt to new tasks and environments with few demonstrations. This limitation significantly hinders the deployment of robots in diverse real-world scenarios, where large amounts of task-specific training data are often impractical or prohibitively expensive.

To tackle this, previous works (Torne et al., 2024; Mandlekar et al., 2023; Jiang et al., 2024; Mu et al., 2024) have focused on augmenting expert demonstrations in simulation to enhance policy learning. However, these approaches typically assume *a priori* knowledge of object and environment CAD models, as well as precise 3D pose estimations — requirements often impractical in real-world scenarios. An alternative line of research (Ma et al., 2023; Li et al., 2024b; Hsu et al., 2024; Zhu et al., 2024) explores learning manipulation directly from visual demonstrations. Some methods leverage human-centric video datasets to learn generalizable representations and generate rewards to learn novel skills (Ma et al., 2023; Li et al., 2024b). Nevertheless, these methods still struggle to learn complex skills, such as deformable and fluid-like manipulations, due to their reliance on physical simulators and the resulting significant sim-to-real gaps. Others acquire manipulation skills from limited human demonstration videos (Hsu et al., 2024; Zhu et al., 2024), yet often rely on off-the-shelf grasping modules and robust pose estimations, limiting scalability in the real world.

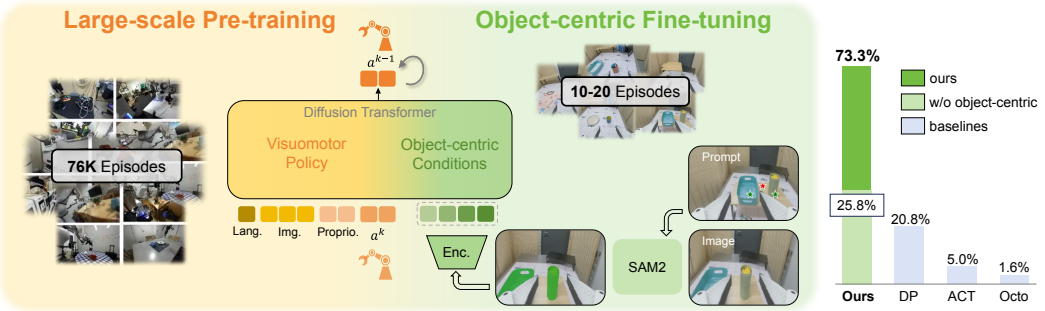

Figure 1: **ControlManip bridges pre-trained manipulation policies with object-centric representations via ControlNet-style efficient fine-tuning.** ControlManip requires only 10–20 demonstrations to achieve 73.3% task success rate, significantly surpassing baseline's 20.8% success rate.

Currently, imitation learning has achieved impressive success in manipulation (Brohan et al., 2023; Fu et al., 2024; Chi et al., 2024; 2023; Liu et al., 2024; Nair et al., 2023; Team et al., 2024; Kim et al., 2024), primarily due to its scalability and capability to acquire skills across a wide range of scenarios without relying on external *a priori* knowledge. In particular, **pre-training general-purpose visuomotor policies** (Brohan et al., 2022; 2023; Kim et al., 2024; Team et al., 2024; Liu et al., 2024; O'Neill et al., 2023) has emerged as a promising approach for enabling generalizable robot behaviors across various tasks and environments. However, fine-tuning these pre-trained policies for downstream tasks remains data-intensive, as substantial amounts of task- and environment-specific data are still required to adapt to the visual and action domains of the target task (Kim et al., 2024; Team et al., 2024; Liu et al., 2024). On a separate front, **object-centric representations** has shown potential in improving data efficiency for learning expert policies (Zhu et al., 2023b;a; Hsu et al., 2024). By focusing on relevant object properties (*e.g.*, shape, size and position), object-centric representations reduce the complexity of the input observation space. This approach enhances policy robustness to changes in object pose and instance, while also making the policies less susceptible to real-world noise compared to pixel-level features. Despite this, existing methods still requires hundreds of demonstrations to learn a task (Zhu et al., 2023b;a). This limitation is largely attributed to the lack of a visuomotor action prior, which is critical for guiding the learning process in data-scarce scenarios.

We introduce **ControlManip**, a novel learning framework that synergistically combines pre-trained visuomotor policies with object-centric representations to enable efficient few-shot learning. By integrating object-centric representations into pre-trained visuomotor policies (Diffusion Transformer in our implementation), our approach leverages both the rich prior knowledge from large-scale pre-training and the data efficiency of object-centric learning. Inspired by Zhang *et al.* (Zhang et al., 2023), ControlManip introduces **additional cross-attention layers with zero-initialized key-value (KV) projection weights**, allowing expert policies to acquire task-specific skills while progressively integrating object-centric representations. This design ensures that the policy focuses on task-relevant concepts without compromising the generalization or action quality of the pre-trained model. The zero-initialization of additional KV projections stabilizes fine-tuning by mitigating introduction of harmful noise, thereby enabling a seamless integration of task-specific object-centric representations with general-purpose visuomotor pre-training. Owing to these advantages, ControlManip significantly reduces data requirements for adapting task-specific policies, enhancing the efficiency of robotic manipulation deployment in the real world.

We demonstrate the efficacy and efficiency of ControlManip across **6 diverse real-world tasks**, achieving robust performance with only **10-20 demonstrations per task**. The evaluation tasks span diverse manipulation challenges: pick-and-place tasks with rigid, soft, and precision-critical objects, as well as complex manipulations including articulated object operation, fluid-like object pouring, and deformable cloth folding. Empirically, ControlManip attains an impressive **73.3%** success rate across all tasks with very limited demonstrations, significantly surpassing baseline methods that achieve a mere 20.8% success rate. Additionally, we also demonstrate the robustness of ControlManip when deployed to unseen objects and backgrounds. Ablation studies confirm the necessity of three key components: (1) visuomotor policy pre-training for skill priors, (2) object-centric representation fusion for task grounding, and (3) ControlNet-style conditioning for stable fine-tuning.

Our primary contribution bridges the gap between **large-scale policy pre-training** and **efficient object-centric adaptation**, enabling robots to acquire complex skills from minimal demonstrations. Beyond advancing few-shot manipulation, ControlManip establishes a blueprint for unifying foundational Vision-Language-Action model (VLA) with structured visual representations — a critical step toward scalable real-world robot learning.

2 Related Works

2.1 Few-shot Learning for Manipulation

Reducing dependence on costly demonstration data while maintaining task performance remains a cornerstone challenge in robotic manipulation. Early approaches focused on *automated data augmentation* (Torne et al., 2024; Mandlekar et al., 2023) by synthesizing large-scale training trajectories in simulation. Complementary work integrates reinforcement learning with imitation learning to improve robustness (Mu et al., 2024). However, these methods demand precise object pose priors and CAD models, hindering deployment in real-world settings where such information is unavailable. Further, reliance on simulated physics engines often creates insurmountable sim-to-real gaps, particularly for deformable objects or contact-rich tasks.

To bypass simulated data, recent efforts leverage *human video priors* to guide robotic policies. Representations distilled from egocentric datasets (*e.g.*, R3M (Nair et al., 2023), VIP (Ma et al., 2023)) encode task-agnostic visual features, enabling skill transfer to robots (Li et al., 2024b). While effective for rigid object manipulation, these methods struggle with precise and dynamic interactions due to their reliance on 2D visual correspondences rather than actionable 3D spatial reasoning.

The frontier of few-shot learning targets direct adaptation from minimal human demonstrations. DenseMatcher (Zhu et al., 2024) localizes 3D semantic correspondences to generalize skills across object instances, while SPOT (Hsu et al., 2024) decouples task planning (via SE(3) object trajectories) from robot actuation. Despite progress, most methods still depend on (1) handcrafted grasping subroutines and (2) accurate pose estimation pipelines, limiting flexibility in unstructured environments. Critically, few explore integration with *pre-trained visuomotor policies* for leveraging large-scale robotic datasets as priors. Our work bridges this gap through ControlNet-style conditioning, enabling efficient few-shot adaptation of general-purpose policies via object-centric representations.

2.2 Object-centric Representation Learning

Object-centric representation learning has attracted significant attention in both robotics and computer vision. By decomposing complex scenes into manipulable objects, these approaches facilitate more efficient reasoning for tasks such as grasping and object manipulation. Traditional methods often represent objects by pose (Tremblay et al., 2018; Tyree et al., 2022; Migimatsu & Bohg, 2020) or bounding boxes (Wang et al., 2019; Devin et al., 2018), which explicitly encode spatial positioning and extent. Although these approaches have proven successful in controlled settings, they frequently rely on prior knowledge of object categories or instance labels, limiting their adaptability to unseen objects and dynamic environments.

To address these issues, unsupervised object discovery methods (Locatello et al., 2020; Burgess et al., 2019) aim to learn representations by autonomously segmenting visual input into meaningful object-like components. However, these methods often encounter challenges in highly cluttered or partially occluded scenes, where objects can overlap significantly, causing poor segmentation or inconsistent object identities (Wang et al., 2021; Heravi et al., 2023). Consequently, their applicability to real-world manipulation, which involves unpredictable object shapes and positions, remains limited.

Much of the prior literature on object-centric representation learning has also struggled to exploit large-scale pre-trained models effectively (Didolkar et al., 2024; Yoon et al., 2023). For example, existing methods (Gao et al., 2023; Yi et al., 2022) typically align representations with category-specific features or pose estimations, creating a mismatch with generic large-scale models trained on extensive, heterogeneous datasets. This incompatibility can necessitate cumbersome, task-specific tuning that undermines the promised benefits of transfer learning, such as faster convergence and deeper semantic understanding.

Despite these limitations, object-centric representation learning holds considerable potential if it can be integrated with large-scale, data-driven representation algorithms. In the subsequent sections, we detail our approach to bridging the gap between object-centric representations and large-scale model pre-training, aiming to overcome these longstanding challenges.

2.3 ControlNet-style Fine-tuning

ControlNet (Zhang et al., 2023) enhances large-scale pre-trained Stable Diffusion (Stability, 2022) by efficiently incorporating additional conditional inputs, such as sketch, normal map, depth map or human pose, through zero-initialized convolution layers. These layers start with zero weights and bias, ensuring no initial impact on outputs while progressively learning to integrate new conditions. This methodology has been extensively applied in various domains, such as controllable image generation (Zhang et al., 2023; Li et al., 2024a; Zavadski et al., 2023), video generation (Guo et al., 2024; Wang et al., 2024a; Bar-Tal et al., 2024), and human motion generation (Dai et al., 2024; Xie et al., 2024). Despite its widespread adoption in these areas, the application of ControlNet in the context of robotic manipulation has not yet been investigated. Our study represents the first effort to adapt ControlNet-style fine-tuning to this field, unifying large-scale visuomotor pre-training with fine-tuning of object-centric representations to enable few-shot robotic manipulation learning.

3 Preliminary

We formulate robot manipulation as a implicit discrete-time Markov Decision Process (MDP) $\mathcal{M} = (\boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{R}}, \gamma)$, where $\boldsymbol{\mathcal{S}}$ represents the state space, $\boldsymbol{\mathcal{A}}$ the action space, $\boldsymbol{\mathcal{T}}$ the stochastic transition probability, $\boldsymbol{\mathcal{R}}$ the binary reward function with $\boldsymbol{r} \in \{0, 1\}$, and $\gamma \in [0, 1)$ the discount factor. In our context, $\boldsymbol{\mathcal{S}}$ denotes the set of all possible configurations of the robot and its environment, while $\boldsymbol{\mathcal{A}}$ denotes the space of the robot's motor commands at each discrete time t. Our objective is to learn a closed-loop visuomotor policy $\boldsymbol{\pi} : \boldsymbol{\mathcal{O}} \rightarrow \boldsymbol{\mathcal{A}}$, where $\boldsymbol{\mathcal{O}}$ is the observation space consisting of the robot proprioception, RGB images and language instruction, which serves as a partial projection of the state space $\boldsymbol{\mathcal{S}}$ derived from the real-world sensors.

Diffusion policy (Chi et al., 2023) formulates the visuomotor policy $\boldsymbol{\pi}$ as the Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2020), which can model complex multimodal action distributions and facilitate a stable training behavior. DDPM performs K iterations of a denoising process, starting from a Gaussian noise $\boldsymbol{x}^K \sim \mathcal{N}(0, I)$ and evolving toward the desired output $\boldsymbol{x}^0 \sim q_\theta\left(\boldsymbol{x}^0\right)$. The denoising process is described by the following equation:

$$\boldsymbol{x}^{k-1} = \alpha\left(\boldsymbol{x}^k - \beta \boldsymbol{\epsilon}_{\boldsymbol{\theta}}\left(\boldsymbol{x}^k, k\right)\right) + \sigma \mathcal{N}(0, I), \tag{1}$$

where α, β, and σ are functions of the timestep k, collectively known as the noise schedule, and the $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$ is the distribution shift prediction network with the trainable parameter θ.

The training objective is to minimize the variational lower bound of KL-divergence between the given data distribution $p\left(\boldsymbol{x}^0\right)$ and the θ-parameterized distribution $q_\theta\left(\boldsymbol{x}^0\right)$. As shown in (Ho et al., 2020), the loss function can be simplified as:

$$\mathcal{L} = \mathbb{E}_{t \sim [1,K], \boldsymbol{x}^0, \epsilon^k}\left[\|\epsilon^k - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}\left(\boldsymbol{x}^0 + \epsilon^k, k\right)\|^2\right]. \tag{2}$$

Diffusion policy represents the robot actions $\boldsymbol{a}_{\boldsymbol{t}:\boldsymbol{t}+\boldsymbol{T_a}}$ as the model output x and conditions the denoising process on the robot observations $\boldsymbol{o}_{\boldsymbol{t}:\boldsymbol{t}-\boldsymbol{T_o}}$, where $\boldsymbol{a_t} \in \boldsymbol{\mathcal{A}}$, $\boldsymbol{o_t} \in \boldsymbol{\mathcal{O}}$, T_a and T_o denote the horizon lengths of the action and observation sequences. For convenience, we use $\boldsymbol{A_t}$ and $\boldsymbol{O_t}$ to represent the action and observation sequences in the following discussion. The DDPM is naturally extended to approximate the conditional distribution $p\left(\boldsymbol{A_t} \mid \boldsymbol{O_t}\right)$ for planning. To capture the conditional actions distribution, the denoising process is modified from Eq. (1):

$$\boldsymbol{A}_{\boldsymbol{t}}^{k-1} = \alpha\left(\boldsymbol{A}_{\boldsymbol{t}}^k - \beta \boldsymbol{\epsilon}_{\boldsymbol{\theta}}\left(\boldsymbol{A}_{\boldsymbol{t}}^k, k\right)\right) + \sigma \mathcal{N}(0, I). \tag{3}$$

The training loss is modified from Eq. (2):

$$\mathcal{L} = \mathbb{E}_{t \sim [1,K], \boldsymbol{A}_{\boldsymbol{t}}^0, \epsilon^k}\left[\|\epsilon^k - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}\left(\boldsymbol{A}_{\boldsymbol{t}}^0 + \epsilon^k, \boldsymbol{O_t}, k\right)\|^2\right]. \tag{4}$$

In practice, we exclude observation features from the denoising process for better accommodates real-time robot control, while the formulation remains the same.

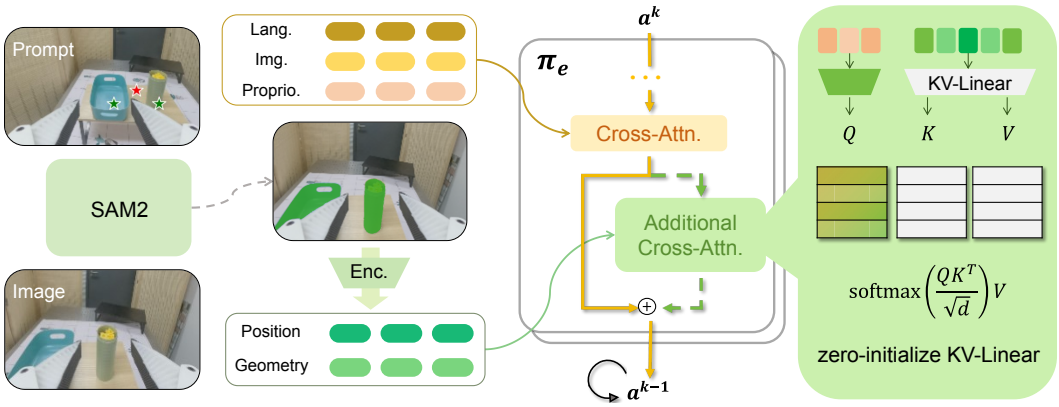

Figure 2: **Overview of ControlManip.** ControlManip leverages a ControlNet-style fine-tuning strategy to integrate object-centric representations with the pre-trained visuomotor policy. The zero-initialized weights and biases preserve the rich prior knowledge of the pre-trained policy while progressively grounding it in structured object properties.

4 METHOD

Given a pre-trained general-purpose policy $\boldsymbol{\pi_g} : \boldsymbol{\mathcal{O}} \rightarrow \boldsymbol{\mathcal{A}}$, our goal is to efficiently learn a task-specific expert policy $\boldsymbol{\pi_e}$ using a limited set of expert demonstrations. To achieve this, we introduce ControlManip, which leverages a ControlNet-style fine-tuning strategy to integrate object-centric representations with the capabilities of the pre-trained visuomotor policy (see Fig. 2). We first pre-train a language-conditioned visuomotor diffusion transformer policy on a large-scale manipulation dataset across diverse tasks (Sec. 4.1). Next, we design an object-centric representation to identify the key concepts for the specific task, thereby focusing the learning process on task-relevant objects (Sec. 4.2). We then explain how ControlNet-style object-centric finetuning allows the policy to acquire task-specific skills efficiently by gradually incorporating object-centric features into the already trained model (Sec. 4.3). Finally, we present crucial implementation details to facilitate reproducibility of our results (Sec. 4.4).

4.1 VISUOMOTOR POLICY PRE-TRAINING

We begin by pre-training a general-purpose policy $\boldsymbol{\pi_g}$, using the public large-scale manipulation datasets $\boldsymbol{\mathcal{D}_g} = \left\{(\boldsymbol{o_t}, \boldsymbol{a_t})_{t=1}^{T_i}\right\}_{i=1}^{N_g}$ across a diverse range of tasks and scenes, where N_g represents the total number of episodes. Formally, we use $\boldsymbol{\pi_g} : \boldsymbol{\mathcal{O}} \rightarrow \boldsymbol{\mathcal{A}}$ to model the conditional action distribution $p\left(\boldsymbol{A_t} \mid \boldsymbol{O_t}\right)$, where $\boldsymbol{A_t}$ represents the future action sequence and $\boldsymbol{O_t}$ denotes the history of the observations. The observation at time t consists of a single-view RGB image $\boldsymbol{I_t}$, a language instruction ℓ_t, and the robot's proprioceptive state $\boldsymbol{q_t}$, such that $\boldsymbol{o_t} = [\boldsymbol{I_t}, \ell_t, \boldsymbol{q_t}]$. The image $\boldsymbol{I_t}$ and language instruction ℓ_t are tokenized via pre-trained encoders and projected into a shared embedding space through a linear projection layer, while the proprioceptive state $\boldsymbol{q_t}$ is similarly embedded using Multi-layer Perceptrons (MLPs).

The $\boldsymbol{\pi_g}$ model adopts a diffusion policy architecture (Chi et al., 2023) to capture the multimodal conditional action distribution. During training, the action sequence is supervised using a conditional denoising loss Eq. (2). At inference, actions are sampled by iteratively denoising started from a pure Gaussian noise $\boldsymbol{A_t^K} \sim \mathcal{N}(0, I)$ into the desired action $\boldsymbol{A_t^0} \sim q_\theta\left(\boldsymbol{A_t^0} \mid \boldsymbol{O_t}\right)$ via Eq. (3), with the process accelerated using Denoising Diffusion Implicit Model (DDIM) (Song et al., 2021) for real-time control.

To effectively integrate the multimodal heterogeneous observations $\boldsymbol{\mathcal{O}_t}$, we choose the Transformer (Vaswani, 2017) as the backbone of the action denoising network $\boldsymbol{\epsilon_\theta}$. The Transformer backbone, with its cross-attention mechanism, enables the policy to efficiently model the complex relations between heterogeneous conditions and outputs, such as visual inputs, language instructions, proprioceptions, and the resulting actions.

4.2 OBJECT-CENTRIC REPRESENTATIONS

This section outlines the process of building object-centric representations $\boldsymbol{Z} \in \boldsymbol{\mathcal{Z}}$ as additional action conditions, which enables task-specific expert policy $\boldsymbol{\pi_e}$ to explicitly identify the key concepts of the task and efficiently learn from few-shot demonstrations. The process involves two key steps: (i) subscribing and tracking task-relevant objects, and (ii) extracting object-centric representations.

Subscribe and track task-relevant objects. To build object representations that consistently attend to task-relevant objects, it is essential to inform the model about their locations and local geometry in the RGB image observation $\boldsymbol{I}$. Our goal is to automatically access fine-grained instance masks $\left\{\boldsymbol{M^i}\right\}_{i=1}^{N_{\text{obj}}}$ corresponding to task-relevant objects $\left\{\mathbf{obj^i}\right\}_{i=1}^{N_{\text{obj}}}$, where N_{obj} denotes the number of objects. Prior works on unsupervised object discovery require extensive training data while being limited to simplified toy domains (Locatello et al., 2020), and rely heavily on third-person viewpoint observation (Zhu et al., 2023b). Inspired by Zhu *et al*. (Zhu et al., 2023a), we adopt a few-shot semi-supervised approach to access task-relevant object instance masks with minimal human expert demonstration. To achieve this, we employ SAM2 (Ravi et al., 2024), a powerful video instance segmentation that processes simple prompts. This allows a demonstrator to subscribe objects prompt in $1-2$ images with just a few mouse clicks by marking keypoints. Notably, a single subscription of objects prompt supports both offline video segmentation on training data, and real-time object tracking during inference.

Learn to extract object-centric representations. We aim to learn a $\boldsymbol{f_\varphi}$ to extract the object-centric representations $\boldsymbol{z^i}$ from $i-\text{th}$ object mask $\boldsymbol{M^i}$ as additional conditions for the expert policy $\boldsymbol{\pi_e}$. To encode *where* and *what* relevant objects are, we design *positional feature* and *geometrical feature* for each object. For positional feature $\boldsymbol{z^i_{\text{pos}}}$, we encode the mean coordinates of the object mask on images using sinusoidal positional encoding (Vaswani, 2017). For geometrical feature $\boldsymbol{z^i_{\text{geo}}}$, we obtain a spatial feature map with a Convolutional Neural Network (CNN) (Krizhevsky et al., 2012) that runs on the mask of each task-relevant object. Similar to the approach of Zhu *et al*. (Zhu et al., 2023b), we train the spatial network from scratch rather than using a pre-trained model, as we require actionable visual features that are specifically informative for continuous control tasks. Finally, we concatenate the positional and geometry feature to form the object-centric representation $\boldsymbol{z^i} = \left[\boldsymbol{z^i_{\text{pos}}}, \boldsymbol{z^i_{\text{geo}}}\right]$ and $\boldsymbol{Z} = \left\{\boldsymbol{z^i}\right\}_{i=1}^{N_{\text{obj}}} \in \boldsymbol{\mathcal{Z}}$

4.3 CONTROLNET-STYLE FINE-TUNING

Given a small set of task-specific dataset $\boldsymbol{\mathcal{D}_e} = \left\{(\boldsymbol{o_t}, \boldsymbol{z_t}, \boldsymbol{a_t})_{t=1}^{T_i}\right\}_{i=1}^{N_e}$, where N_e represents the number of demonstrations, we aim to efficiently fine-tune an expert policy $\boldsymbol{\pi_e} : \boldsymbol{\mathcal{O}} \times \boldsymbol{\mathcal{Z}} \rightarrow \boldsymbol{\mathcal{A}}$ from $\boldsymbol{\pi_g} : \boldsymbol{\mathcal{O}} \rightarrow \boldsymbol{\mathcal{A}}$ with the object-centric representations.

In our context, the pre-trained policy is a transformer-based architecture that utilizes cross-attention blocks to model actions conditioned on observations. Specifically, the cross-attention mechanism computes the relationship between actions $\boldsymbol{A}$ and observations $\boldsymbol{O}$ as:

$$\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d}}\right)\boldsymbol{V}, \tag{5}$$

where $\boldsymbol{Q} = \mathbf{W_a}\boldsymbol{A} + \mathbf{B_a}$ represents the query projection, and $\boldsymbol{K}, \boldsymbol{V} = \mathbf{W_o}\boldsymbol{O} + \mathbf{B_o}$ represent the key and value projections, respectively. To incorporate the object-centric representation $\boldsymbol{Z} \in \boldsymbol{\mathcal{Z}}$, we extend the cross-attention mechanism by introducing a dual-attention structure. Instead of directly appending $\boldsymbol{Z}$ as additional tokens to $\boldsymbol{K}$ and $\boldsymbol{V}$, we compute a separate attention branch for the object-centric conditions:

$$\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d}}\right)\boldsymbol{V} + \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K_z}^T}{\sqrt{d}}\right)\boldsymbol{V_z}, \tag{6}$$

where $\boldsymbol{K_z}, \boldsymbol{V_z} = \mathbf{W_z}\boldsymbol{Z} + \mathbf{B_z}$ are the key and value projections for the object-centric observations $\boldsymbol{Z}$.

Inspired by Zhang *et al*. (Zhang et al., 2023), we *zero-initialize* the additional KV-projection layers to ensure the expert policy $\boldsymbol{\pi_e}$ model behaves similarly to the pre-trained general-purpose policy $\boldsymbol{\pi_g}$

during the early stage of fine-tuning, preserving the model's prior knowledge. Since the weight $\mathbf{W_z}$ and bias $\mathbf{B_z}$ are initialized to $\mathbf{0}$, the key and value projections for $\boldsymbol{Z}$ are zero:

$$\mathbf{K_z} = \mathbf{W_z}\boldsymbol{Z} + \mathbf{B_z} = \mathbf{0}, \quad \mathbf{V_z} = \mathbf{0}. \tag{7}$$

Thus, the dual-attention output in Eq. (6) reduces to the original cross-attention in Eq. (5), preserving the pre-trained policy's behavior at the first fine-tuning step.

A common misunderstanding with zero-initialized weights and biases is that they produce zero gradients and are, therefore, untrainable. We demonstrate that the additional KV-projection layers $(\mathbf{W_z}, \mathbf{B_z})$ and the object-centric representations $\boldsymbol{Z}$ can be optimized despite their zero initialization, which is similar to the case in ControlNet (Zhang et al., 2023).

Let $\frac{\partial \mathcal{L}}{\partial \mathbf{V_z}}$ denote the upstream gradient from the loss $\mathcal{L}$. The gradients for $\mathbf{W_z}$ and $\mathbf{B_z}$ are:

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial \mathbf{W_z}} = \displaystyle\sum_{p,i} \frac{\partial \mathcal{L}}{\partial \boldsymbol{V_z}_{p,i}} \cdot \boldsymbol{Z}_{p,i} \\ \dfrac{\partial \mathcal{L}}{\partial \mathbf{B_z}} = \displaystyle\sum_{p,i} \frac{\partial \mathcal{L}}{\partial \boldsymbol{V_z}_{p,i}} \cdot 1 \end{cases} \tag{8}$$

Since $\boldsymbol{Z}$ is non-zero, $\frac{\partial \mathcal{L}}{\partial \mathbf{W_z}} \neq \mathbf{0}$ if $\frac{\partial \mathcal{L}}{\partial \boldsymbol{V_z}} \neq \mathbf{0}$. Similarly, $\frac{\partial \mathcal{L}}{\partial \mathbf{B_z}}$ accumulates non-zero gradients. After one gradient step:

$$\begin{cases} \mathbf{W_z}^* = \mathbf{W_z} - \beta_l \cdot \dfrac{\partial \mathcal{L}}{\partial \mathbf{W_z}} \neq \mathbf{0} \\ \mathbf{B_z}^* = \mathbf{B_z} - \beta_l \cdot \dfrac{\partial \mathcal{L}}{\partial \mathbf{B_z}} \neq \mathbf{0} \end{cases} \tag{9}$$

This ensures $\boldsymbol{K_z^*}$ and $\boldsymbol{V_z^*}$ become non-zero, allowing the dual-attention to incorporate $\boldsymbol{Z}$.

Considering $\boldsymbol{Z}$ is learnable, its gradient is:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{Z}} = \mathbf{W_z}^T \cdot \frac{\partial \mathcal{L}}{\partial \boldsymbol{V_z}}. \tag{10}$$

Since $\mathbf{W_z}^* \neq \mathbf{0}$, $\boldsymbol{Z}$ receives non-zero gradients and is updated accordingly. This aligns with the zero-convolution principle, where gradients persist despite zero-initialized parameters. We fine-tune the expert policy using the conditional denoising loss as defined in Eq. (4).

With the ControlNet-style fine-tuning, we efficiently integrate additional object-centric conditions into the pre-trained visuomotor policy. This approach ensures that when the KV-projection layers are zero-initialized in the dual cross-attention module, the deep neural features remain unaffected prior to any optimization. The capabilities, functionality, and output action quality of the pre-trained visuomotor modules are perfectly preserved, while further optimization becomes as efficient as standard fine-tuning. This allows ControlManip to simultaneously leverage the advantages of large-scale pre-training and object-centric representations, accelerating real-world robot adoption by significantly reducing the data requirements for task deployment.

4.4 Implementation Details

In Sec. 4.1, we pre-train the policy $\boldsymbol{\pi_g}$ on the full DROID dataset (Khazatsky et al., 2024), using the wrist camera image $\boldsymbol{I_t}$, end-effector poses and gripper widths $\boldsymbol{q_t}$, and episode language descriptions ℓ_t. The observation and action horizons are set to $T_o = 2$ and $T_a = 16$. The pre-trained policy, implemented as a Diffusion Transformer (Chi et al., 2023) with 29M parameters, uses a CLIP (Radford et al., 2021) ViT-B/16 vision encoder and a Transformer text encoder. We pre-train $\boldsymbol{\pi_g}$ with AdamW (learning rates: 1×10^{-4} for denoising model, 3×10^{-5} for vision; text encoder frozen) on 4 NVIDIA A800 GPUs for 3 days. In Sec. 4.2, we extract object-centric representations from raw images. In Sec. 4.3, we fine-tune $\boldsymbol{\pi_e}$ on evaluation tasks, adding $\sim$5M parameters. Fine-tuning uses the same settings as pre-training and runs on a single NVIDIA A800 GPU for 12 hours.

5 Experiments

To evaluate the efficiency of our proposed method, we conduct 6 various real-world tasks, utilizing only 10-20 demonstrations. Empirically, our findings indicate that our method consistently and

Table 1: **Illustrations of Evaluation Tasks.** We develop a suite of 6 real-world tasks for evaluation, including pick-and-place various types of object (*e.g.*, rigid, soft, and precise), as well as articulated, deformable and fluid-like manipulations.

TASK NAME	TASK TYPE	#DEMOS	LANGUAGE DESCRIPTION
`RearrangeCup`	Rigid Pick&Place	14	Rearrange the cup and set it on the light plate.
`OrganizeToy`	Soft Pick&Place	20	Pick up the green toy and place it in the blue bowl.
`OrganizeScissors`	Precise Pick&Place	15	Pick up the scissors from the pen holder into the blue basket.
`OpenCabinet`	Articulated Manipulation	11	Open the cabinet with the black handle.
`FoldClothes`	Deformable Manipulation	16	Fold up the sleeves of the pink clothing on the table.
`PourCubes`	Pouring Behavior	19	Pour the blocks from the green cup into the blue box.

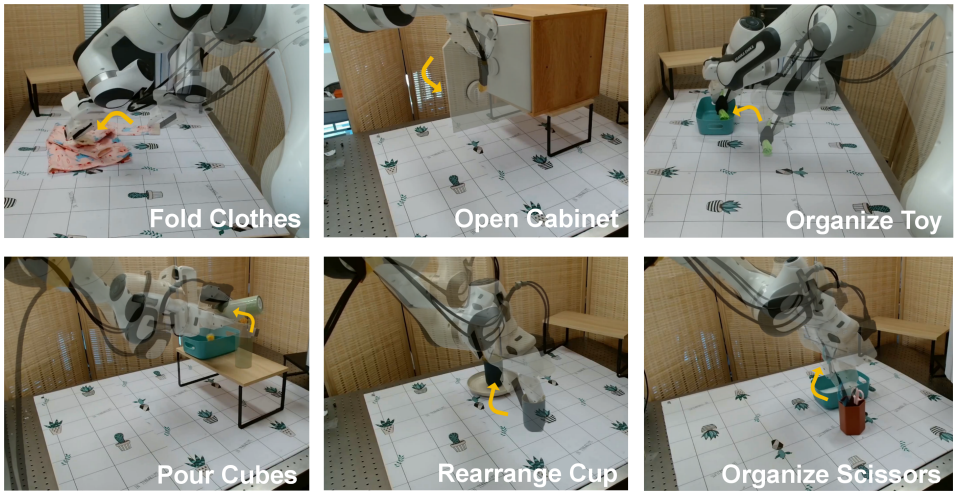

Figure 3: **Task Visualization.** We show the evaluation setup for the six evaluation tasks. The initial and target states are shown as transparent and solid layers, respectively. The yellow arrow highlights the desired transition.

significantly improves success rates across all tasks, achieving an overall success rate of **73.3%**, which markedly surpasses the baseline of 20.8%. We further evaluate the policies performance on data scaling. Our results show that our method rapidly converges to a high success rate with as few as 20 demonstrations, while baselines require more than 100 demonstrations to achieve comparable performance. Additionally, we also demonstrate the robustness and generalization of our method over unseen objects and backgrounds.

5.1 Experimental Setup

Tasks. We develop a suite of 6 various real-world tasks to evaluate the efficacy of our proposed method. These tasks are designed to cover a wide range of manipulation challenges, including pick-and-place various types of objects like rigid `RearrangeCup`, soft `OrganizeToy`, and precise `OrganizeScissors`, as well as articulated `OpenCabinet`, deformable `FoldClothes`, and fluid-like `PourCubes` manipulations. A detailed illustration of the dimensions and definitions of each task is given in Tab. 1 while visualization is provided in Fig. 3.

Data Collection. We collect a small set of demonstrations for each evaluation task with UMI (Chi et al., 2024), an arm-agnostic data collection system equipped with a hand-held gripper to efficiently gather demonstrations. UMI is equipped with a wrist-mounted GoPro camera, which provides access to the RGB image and relative end-effector 6D pose trajectory through visual SLAM. The number of demonstrations for each evaluation task is detailed in Tab. 1. In each demonstration, the positions of the objects and the hand-held UMI device are randomly initialized.

Baselines and Ablations. We compare our method against Octo (Team et al., 2024), VIOLA (Zhu et al., 2023b), ACT (Zhao et al., 2023), and Diffusion Policy (Chi et al., 2023). Octo is a pre-trained foundation VLA on the large-scale RT-X dataset (O'Neill et al., 2023), employing a diffusion-based model to decode the action tokens. VIOLA is a widely recognized 2D object-centric transformer-

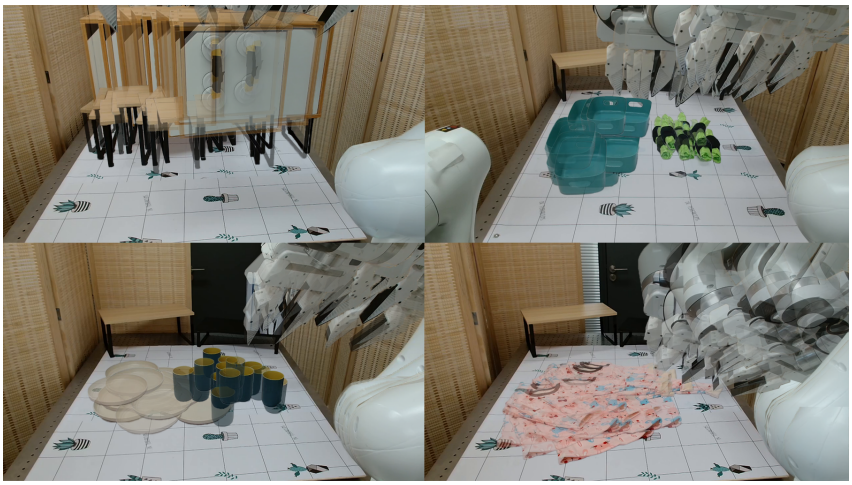

Figure 4: **Initial state distribution of policy evaluation.** We show the initial objects and robot states distribution for policy evaluation on `OpenCabinet`, `OrganizeToy`, `RearrangeCup` and `FoldClothes`. The initial states follow the same distribution as the training episodes.

based policy learning framework that utilizes Region Proposal Network (RPN) (Zhou et al., 2022) to extract object-centric representations. ACT and Diffusion Policy are among the most extensively studied and widely applied imitation visuomotor policies. ACT models the actions using a Variational Autoencoder (VAE), while Diffusion Policy leverages DDPM to capture more multi-modal action distributions. For a fair comparison, we implement Diffusion Policy using the same architecture as our base model, excluding the object-centric modules.

In our ablation study, we systematically remove individual components from our method to investigate their independent contributions. We ablate the pre-training phase by training an object-centric Diffusion Policy from scratch, denoted as "w/o pre-train". We eliminate the object-centric representations by directly fine-tuning the pre-trained model, denoted as "w/o object-centric". To assess the significance of ControlNet-style fine-tuning in integrating object-centric representations into pre-trained policies, we omit the zero-initialization of the projection layers for additional object-centric conditions, denoted as "w/o zero-init".

Evaluation Setup and Protocol. We deploy a Franka Emika FR3 robotic arm and a Panda gripper, equipped with the same GoPro camera used for data collection, for inference of all policies. Task success rate is used as the main evaluation metric, and each trial is terminated if the policy shows no trend of success, the robot enters a potentially unsafe interaction state with the environment, or the task is completed. In the main experiments, all tasks are evaluated in the same environment as data collection but with randomized initial states for both the robot and objects, as shown in Fig. 4.

5.2 Comparative and Ablation Results

For comparison and ablation studies, we evaluate each model over 20 trials per task. As illustrated in Fig. 5, the task success rates are presented within and across all evaluation tasks, providing a comprehensive overview of our findings.

Our method, ControlManip, achieves an impressive **overall task success rate of 73.3%**, significantly outperforming the baselines. The strongest baseline, Diffusion Policy, attains only a 20.8% success rate and struggles to precisely manipulate target objects or learn complex behaviors such as pouring. Octo and ACT only achieve $2/20$ and $6/20$ success on the `OpenCabinet` task, with no successes in other tasks, resulting in the overall success rate of just 1.6% and 5.0%. Despite Octo's advantage from large-scale pre-training, its regression-based backbone fails to model action distributions with multiple distinct modes. While ACT leverages a VAE to represent action diversity, it is hindered by posterior collapse, especially in the low-data regime. VIOLA, reliant on static workspace cameras for object-centric representations (unavailable in our setup), shows no successful cases. Its object proposal mechanism degrades severely with wrist-camera inputs, highlighting

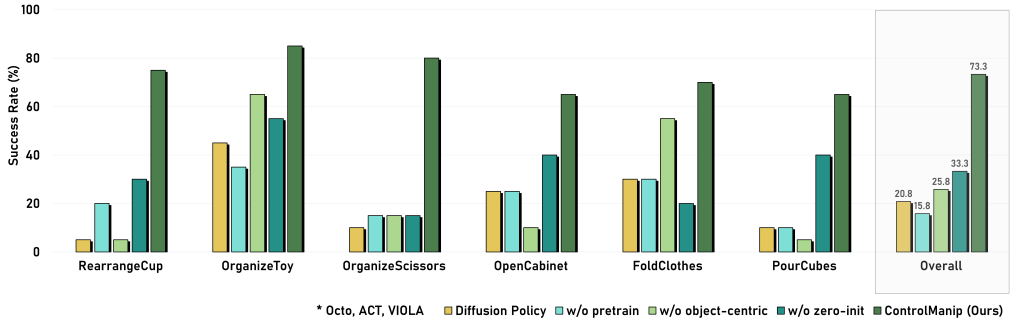

Figure 5: **Main Comparison and Ablation Study.** All policies are trained or fine-tuned from a shared, limited demonstration dataset for each task. *Octo, ACT, and VIOLA are omitted due to very low success rates, with overall success rates of 1.6%, 5.0%, and 0.0%, respectively.

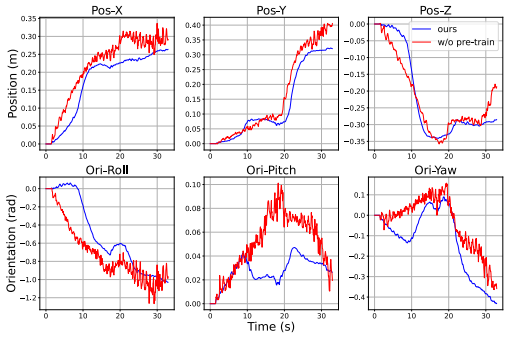

Figure 6: **Jitter Problem Visualization.** End-effector trajectories of our method vs. "w/o pre-train" on `RearrangeCup`. The ablation shows significant jitter across all action dimensions compared to our full model.

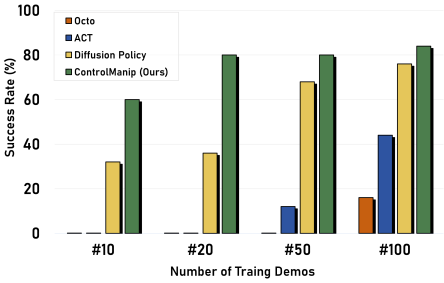

Figure 7: **Performance on Data Scaling.** In the `OrganizeToy` task, ControlManip achieves a high success rate of 80% with 20 demonstrations, while the best baseline requires 100 to reach a comparable performance.

sensitivity to viewpoint. In contrast, ControlManip demonstrates robust and efficient learning across various real-world tasks even when only limited demonstrations are available.

Ablation studies reveal the critical role of key components of ControlManip. Removing the pre-training phase and directly incorporating object-centric representations ("w/o pre-train") results in severe jitter problems when policies deploy to a real robot, as shown in Fig. 6. This instability persists across the `RearrangeCup`, `OrganizeToy`, `OpenCabinet`, and `PourCubes` tasks, significantly reducing success rates — even falling below the Diffusion Policy baseline that relies solely on raw image observations. We hypothesize that the object-centric features may provide deceptive low-loss pathways during the early training stage, incentivizing the policy to bypass learning from the more stable visual features. Eliminating object-centric representations during fine-tuning ("w/o object-centric") provides only a marginal 5.0% improvement over training Diffusion Policy from scratch, highlighting the importance of structured representations. Finally, removing zero-initialization for additional object-centric conditions ("w/o zero-init") causes a drastic 50% drop in success rate, emphasizing the role of proper initialization in stabilizing training and improving task performance. Notably, our complete methodology degenerates to the Diffusion Policy baseline when stripping all three components (pre-training, object-centric representations, and zero-initialization).

5.3 Performance on Data Scaling

We evaluate ControlManip's data efficiency through controlled scaling experiments on the `OrganizeToy` task, benchmarking against established baseline methods. Each approach is tested

across demonstration set sizes of 10, 20, 50, and 100 episodes with 25 trials per condition, with results shown in Fig. 7. While all methods improve as the amount of training data increases, ControlManip achieves a high success rate of 80% with only 20 demonstrations — a level unattained by baselines even at 100 demonstrations. This highlights the efficiency of ControlManip in learning from limited data.

5.4 Generalization on Object and Background

To evaluate generalization and robustness, we tested ControlManip on the `RearrangeCup` task across different objects and backgrounds. After training with 14 demonstrations on a single cup and background, ControlManip achieved a 46.7% success rate (14/30 trials) on three unseen cups and a novel background, requiring only a cup prompt for object mask extraction. While this performance is lower than the 75% in-domain success rate, it demonstrates the potential of ControlManip to adapt to dynamic and diverse environments, highlighting its capability to generalize beyond the training domain.

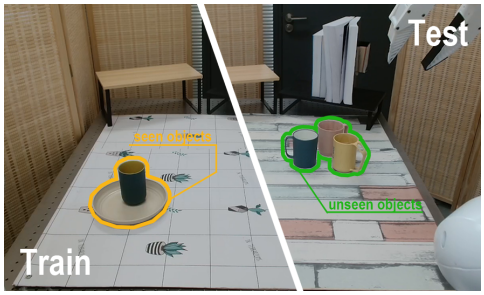

Figure 8: **Generalization over object and background appearance changes.**

6 Limitations and Future Works

While ControlManip demonstrates its efficacy and efficiency across a wide range of manipulation tasks, a few limitations remain, presenting opportunities for future improvements. First, generating object prompts still requires expert human knowledge, as users must manually specify prompts for segmentation. Although this process can be completed with a few simple mouse clicks, it introduces human bias and limits scalability. A promising direction for future work is automating this process through foundational Vision-Language Models (VLMs), allowing segmentation models like SAM2 to generate reasonable object prompts autonomously.

Second, our evaluation tasks are currently constrained to single-arm, short-horizon scenarios, with experiments conducted primarily in controlled laboratory environments. While these settings provide a reliable testbed, they do not fully capture the challenges of real-world deployment. Addressing bi-manual, long-horizon, and in-the-wild manipulation tasks is crucial for improving the generalization of ControlManip. Nevertheless, ControlManip is a general framework, and future work could explore larger pre-trained bi-manual foundation policies and more diverse, efficient object-centric representations to tackle these challenges.

7 Conclusion

This work introduces ControlManip, a framework that bridges large-scale visuomotor policy pre-training with object-centric representations to enable efficient few-shot adaptation for robotic manipulation. By integrating a ControlNet-style architecture, ControlManip injects task-specific object-centric conditions into a pre-trained Diffusion Transformer policy through zero-initialized key-value projection layers. This design preserves the rich prior knowledge of the base policy while progressively grounding it in structured object properties, achieving stable fine-tuning with minimal demonstrations. Across six diverse real-world tasks—ranging from rigid object pick-and-place to deformable cloth folding and fluid-like pouring—ControlManip achieves a 73.3% success rate with only 10–20 demonstrations, outperforming the 20.8% baseline by approximately 252%.

By reducing demonstration requirements to practical levels, ControlManip lowers barriers to deploying robots in diverse scenarios. Overall, our results establish a promising direction for combining large-scale visuomotor priors with structured inputs, setting the stage for scalable few-shot learning and accelerating real-world robot adoption.

ACKNOWLEDGMENTS

We thank Yuyang Li (BIGAI, PKU) for his technical support and contributed discussion, Yuwei Guo (CUHK) for his discussion, and Ziyuan Jiao (BIGAI) for his assistance in setting up the real-world environment.

REFERENCES

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

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.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. Monet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019.

Tianxing Chen, Yao Mu, Zhixuan Liang, Zanxin Chen, Shijia Peng, Qiangyu Chen, Mingkun Xu, Ruizhen Hu, Hongyuan Zhang, Xuelong Li, et al. G3flow: Generative 3d semantic flow for pose-aware and generalizable object manipulation. *arXiv preprint arXiv:2411.18369*, 2024.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, pp. 02783649241273668, 2023.

Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. *arXiv preprint arXiv:2402.10329*, 2024.

Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. Motionlcm: Real-time controllable motion generation via latent consistency model. In *European Conference on Computer Vision*, pp. 390–408. Springer, 2024.

Coline Devin, Pieter Abbeel, Trevor Darrell, and Sergey Levine. Deep object-centric representations for generalizable robot learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7111–7118. IEEE, 2018.

Aniket Didolkar, Andrii Zadaianchuk, Anirudh Goyal, Mike Mozer, Yoshua Bengio, Georg Martius, and Maximilian Seitzer. Zero-shot object-centric representation learning. *arXiv preprint arXiv:2408.09162*, 2024.

Zipeng Fu, Tony Z Zhao, and Chelsea Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. *arXiv preprint arXiv:2401.02117*, 2024.

Ning Gao, Vien Anh Ngo, Hanna Ziesche, and Gerhard Neumann. Sa6d: Self-adaptive few-shot 6d pose estimator for novel and occluded objects. In *7th Annual Conference on Robot Learning*, 2023.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *The Twelfth International Conference on Learning Representations*, 2024.

Negin Heravi, Ayzaan Wahid, Corey Lynch, Pete Florence, Travis Armstrong, Jonathan Tompson, Pierre Sermanet, Jeannette Bohg, and Debidatta Dwibedi. Visuomotor control in multi-object scenes using object-aware representations. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9515–9522. IEEE, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Cheng-Chun Hsu, Bowen Wen, Jie Xu, Yashraj Narang, Xiaolong Wang, Yuke Zhu, Joydeep Biswas, and Stan Birchfield. Spot: Se (3) pose trajectory diffusion for object-centric manipulation. *arXiv preprint arXiv:2411.00965*, 2024.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 20413–20451, 2024.

Zhenyu Jiang, Yuqi Xie, Kevin Lin, Zhenjia Xu, Weikang Wan, Ajay Mandlekar, Linxi Fan, and Yuke Zhu. Dexmimicgen: Automated data generation for bimanual dexterous manipulation via imitation learning. *arXiv preprint arXiv:2410.24185*, 2024.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. In *Robotics: Science and Systems*, 2024.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Kailin Li, Puhao Li, Tengyu Liu, Yuyang Li, and Siyuan Huang. Maniptrans: Efficient dexterous bimanual manipulation transfer via residual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2025.

Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet++: Improving conditional controls with efficient consistency feedback: Project page: liming-ai. github. io/controlnet_plus_plus. In *European Conference on Computer Vision*, pp. 129–147. Springer, 2024a.

Puhao Li, Tengyu Liu, Yuyang Li, Yiran Geng, Yixin Zhu, Yaodong Yang, and Siyuan Huang. Gendexgrasp: Generalizable dexterous grasping. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8068–8074. IEEE, 2023.

Puhao Li, Tengyu Liu, Yuyang Li, Muzhi Han, Haoran Geng, Shu Wang, Yixin Zhu, Song-Chun Zhu, and Siyuan Huang. Ag2manip: Learning novel manipulation skills with agent-agnostic visual and action representations. *arXiv preprint arXiv:2404.17521*, 2024b.

Yuyang Li, Bo Liu, Yiran Geng, Puhao Li, Yaodong Yang, Yixin Zhu, Tengyu Liu, and Siyuan Huang. Grasp multiple objects with one hand. *IEEE Robotics and Automation Letters*, 2024c.

Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33:11525–11538, 2020.

Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. In *The Eleventh International Conference on Learning Representations*, 2023.

Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. Mimicgen: A data generation system for scalable robot learning using human demonstrations. *arXiv preprint arXiv:2310.17596*, 2023.

Toki Migimatsu and Jeannette Bohg. Object-centric task and motion planning in dynamic environments. *IEEE Robotics and Automation Letters*, 5(2):844–851, 2020.

Yao Mu, Tianxing Chen, Shijia Peng, Zanxin Chen, Zeyu Gao, Yude Zou, Lunkai Lin, Zhiqiang Xie, and Ping Luo. Robotwin: Dual-arm robot benchmark with generative digital twins (early version). *arXiv preprint arXiv:2409.02920*, 2024.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. In *Conference on Robot Learning*, pp. 892–909. PMLR, 2023.

Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.

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*, pp. 8748–8763. PMLR, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Stability. Stable diffusion v1.5 model card. `https://huggingface.co/runwayml/stable-diffusion-v1-5`, 2022. Accessed: 2024-09-05.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

Marcel Torne, Anthony Simeonov, Zechu Li, April Chan, Tao Chen, Abhishek Gupta, and Pulkit Agrawal. Reconciling reality through simulation: A real-to-sim-to-real approach for robust manipulation. *arXiv preprint arXiv:2403.03949*, 2024.

Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stan Birchfield. Deep object pose estimation for semantic robotic grasping of household objects. *arXiv preprint arXiv:1809.10790*, 2018.

Stephen Tyree, Jonathan Tremblay, Thang To, Jia Cheng, Terry Mosier, Jeffrey Smith, and Stan Birchfield. 6-dof pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 13081–13088. IEEE, 2022.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Chen Wang, Rui Wang, Ajay Mandlekar, Li Fei-Fei, Silvio Savarese, and Danfei Xu. Generalization through hand-eye coordination: An action space for learning spatially-invariant visuomotor control. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8913–8920. IEEE, 2021.

Dequan Wang, Coline Devin, Qi-Zhi Cai, Fisher Yu, and Trevor Darrell. Deep object-centric policies for autonomous driving. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 8853–8859. IEEE, 2019.

Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for realistic human dance generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9326–9336, 2024a.

Yixuan Wang, Guang Yin, Binghao Huang, Tarik Kelestemur, Jiuguang Wang, and Yunzhu Li. Gendp: 3d semantic fields for category-level generalizable diffusion policy. In *8th Annual Conference on Robot Learning*, volume 2, 2024b.

Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. In *The Twelfth International Conference on Learning Representations*, 2024.

Sherry Yang, Yilun Du, Seyed Kamyar Seyed Ghasemipour, Jonathan Tompson, Leslie Pack Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *The Twelfth International Conference on Learning Representations*, 2024.

Qi Yi, Rui Zhang, Jiaming Guo, Xing Hu, Zidong Du, Qi Guo, Yunji Chen, et al. Object-category aware reinforcement learning. *Advances in Neural Information Processing Systems*, 35:36453–36465, 2022.

Jaesik Yoon, Yi-Fu Wu, Heechul Bae, and Sungjin Ahn. An investigation into pre-training object-centric representations for reinforcement learning. *arXiv preprint arXiv:2302.04419*, 2023.

Denis Zavadski, Johann-Friedrich Feiden, and Carsten Rother. Controlnet-xs: Designing an efficient and effective architecture for controlling text-to-image diffusion models. *arXiv preprint arXiv:2312.06573*, 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pp. 350–368. Springer, 2022.

Junzhe Zhu, Yuanchen Ju, Junyi Zhang, Muhan Wang, Zhecheng Yuan, Kaizhe Hu, and Huazhe Xu. Densematcher: Learning 3d semantic correspondence for category-level manipulation from a single demo. *arXiv preprint arXiv:2412.05268*, 2024.

Yifeng Zhu, Zhenyu Jiang, Peter Stone, and Yuke Zhu. Learning generalizable manipulation policies with object-centric 3d representations. In *Conference on Robot Learning*, pp. 3418–3433. PMLR, 2023a.

Yifeng Zhu, Abhishek Joshi, Peter Stone, and Yuke Zhu. Viola: Imitation learning for vision-based manipulation with object proposal priors. In *Conference on Robot Learning*, pp. 1199–1210. PMLR, 2023b.