Hessisches Kompetenzzentrum
für Hochleistungsrechnen

# *Parameter-Efficient Tuning of Pretrained Visual-Language Models in Multitask Robot Learning*

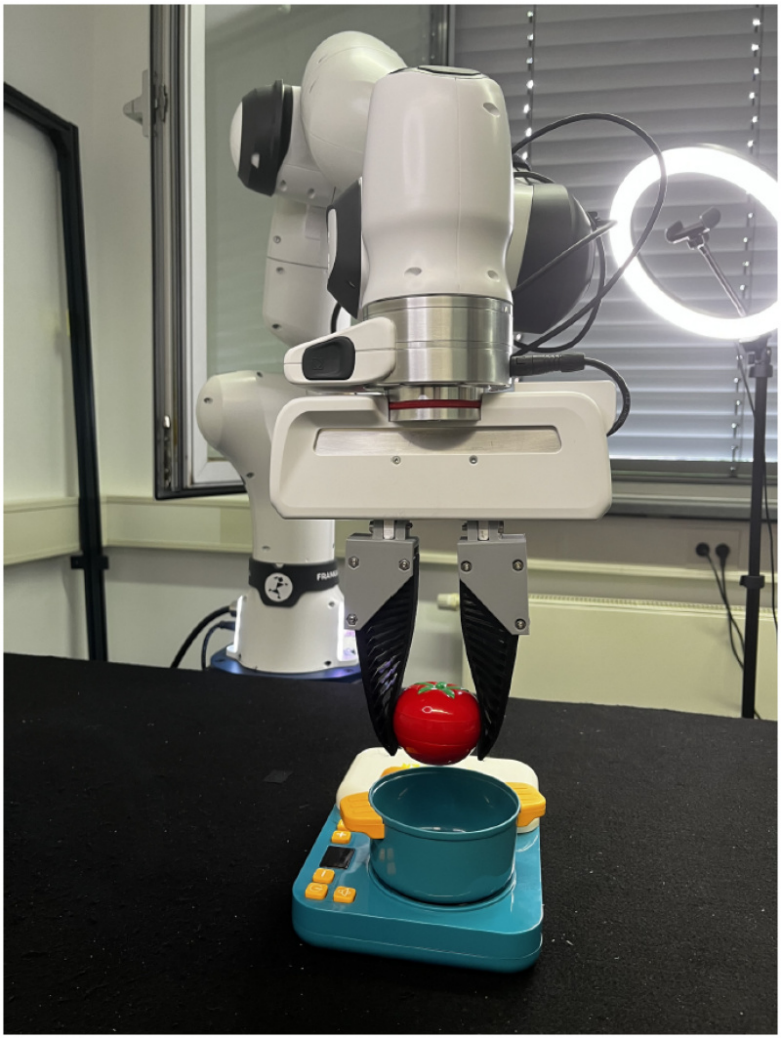

Figure 1: Real-world setup for transfer learning on a Franka Emika robotic arm. In this case, it autonomously performs the language-conditioned task: Pick up the tomato and place it in the pot.

Project Manager
Marcel Mittenbühler

Researchers
Prof. Dr. Carlo D'Eramo, Ahmed Hendawy and Alina Böhm

Principal Investigator
Prof. Dr. Georgia Chalvatzaki

Project Term
2023 - 2024

Clusters
Lichtenberg II Cluster Darmstadt

Software
PyTorch

Institute
Interactive Robot Perception & Learning, Intelligent Autonomous Systems

University
Technische Universität Darmstadt



JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN
TECHNISCHE UNIVERSITÄT DARMSTADT
UNIKASSEL VERSITÄT
Philipps Universität Marburg

## Introduction

General-purpose intelligent robots are expected to simultaneously handle multiple tasks while interpreting various modalities. To this end, efforts to develop multimodal multitask robot policies are trying to keep up with the current success of large visual-language models, which demonstrated state-of-the-art performance in a wide range of tasks at the cost of consuming immense internet-wide data. These attempts aim to leverage the ability of multimodal information to alleviate the need for complicated engineered goal specifications in robotic tasks, offering an excellent opportunity for intuitive interaction with humans, such as through natural language and images. Training state-of-the-art visual language models requires considerable amounts of memory, even when working with parameter-efficient approaches. Moreover, ablations are needed to understand the benefits and limitations of an approach. Access to a compute cluster is essential to execute these experiments due to available memory and the ability to run multiple experiments in parallel.

## Methods

In this work, we propose an alternative approach for enabling efficient learning of multimodal multitask policies aiming to leverage as much prior information as possible. First, we investigated a novel encoder-decoder multimodal temporal diffusion transformer model that affords adaptation. Our encoder merges tokens from pre-trained visual-language models that we adapt to robot-related tasks with the tokens from the robot's proprioception. Those tokens form a dense representation that we pass through our temporal diffusion transformer decoder model which performs temporal attention over the tokens to output robot actions, conditioned on the multimodal observations. Next, we investigated a new adaptation scheme for enabling transferable diffusion policy learning. We showed how to use, where to add, and which types of adapters are necessary for parameter-efficient finetuning of multimodal multitask diffusion policies. Our adapters are composable and can be combined sequentially to perform transfer learning or in parallel to capture more intricate observations-action relations. Unlike finetuning or training from scratch, our approach requires much less data while retaining performance in the originally trained tasks.

## Results

For the adaptation of the encoders, we showed that adapters are a better approach than fine-tuning all parameters in terms of final success rates in manipulation tasks and memory efficiency during training. In this regard, we also validated the results of SpawnNet feature extractors for adaptation. Then, we pre-train a policy on Libero-90– a robotics dataset with 90 language conditioned tasks – as a basis for performing the transfer. Our results suggest that adaptation methods outperform training new policies in light of a few demonstrations. On the flip side, scratch-trained policies outperformed transfer learning once

sufficient demonstrations were available. Finally, we demonstrated adapters as a parameter-efficient approach to bridge the sim-to-real gap with better results than fine-tuning a pre-trained policy. Our experimental results in sim-to-sim transfer show an improvement of up to 31% compared to training from scratch and 5% w.r.t. end-to-end finetuning while using much fewer parameters. Notably, our sim-to-real adaptation of our multimodal diffusion policy for pick and place tasks led to a total success of 90% compared to 85% in full finetuning, with less than one-third of the parameters.

## Discussion

We introduced a novel approach for learning multimodal multitask diffusion policies that can be built on top of large pre-trained visual-language models and that leverages an encoder-decoder transformer architecture for fast adaptation to out-of-distribution tasks. We introduced a composable adaptation scheme that allows fast transfer of diffusion policies even with a handful of demonstrations and transferring knowledge from sim-tosim and sim-to-real tasks, while avoiding forgetting. Crucially, our adaptation scheme is flexible and can be switched-on/off according to the setting, paving the way for the democratization of large pre-trained models in robotics. Our model would still suffer from the common Behavioral Cloning issues, e.g., bad performance in out-of-distribution cases. Future works could investigate the integration of reinforcement learning to resolve this. While we found it necessary to use Bottleneck and LoRA adapters in the diffusion transformer decoder, we did not conclude a universal parameterization that works the best in all cases. A more thorough ablation is needed in this case. Finally, we deem it essential to conduct an explainability analysis to find out how the adapters affect the attention of the network in new settings.

## *Last Update:* 2024-12-10 10:30