OKAMI: Teaching Humanoid Robots Manipulation Skills through Single Video Imitation

Anonymous Author(s)
Affiliation
Address
email

Abstract: We study the problem of teaching humanoid robots to imitate manipulation skills by watching single human videos. To tackle this problem, we investigate an object-aware retargeting approach, where humanoid robots mimic the human motions in the video while adapting to the object locations during deployment. We introduce OKAMI, an algorithm that generates a reference plan from a single RGB-D video, and derive a policy that follows the plan to complete the task. OKAMI sheds light on deploying humanoid robots in everyday environments, where the humanoid robot will quickly adapt to a new task given a single human video. Our experiments show that OKAMI outperforms the baseline by 58.33%, while showcasing systematic generalization across varying visual and spatial conditions. More videos can be found in supplementary materials and website https://sites.google.com/view/okami-corl2024.

Keywords: Humanoid Manipulation, Imitation From Videos, Motion Retargeting

Figure 1: OKAMI enables a human user to teach the humanoid robot how to perform a task by providing a single video demonstration.

1 Introduction

Deploying generalist robots, such as robot butlers, to help with everyday tasks requires them to operate in our daily environments. Humanoid robots, with their human-like embodiment, naturally fit into environments tailored to humans. With recent advancements in hardware design and increased commercial availability, humanoids stand out as an ideal choice for deployment in our living and working spaces. Despite their great potential, humanoid robots struggle to interact

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

autonomously with objects. Recent works have developed deep imitation learning methods for humanoid manipulation [1–3]. However, they require collecting demonstrations through whole-body teleoperation, demanding both expertise and significant physical efforts. In contrast, humans have the ability to watch their partners do a task once and mimic it afterward. Motivated by this observation, we explore the idea of teaching humanoid robots to manipulate objects by watching humans. We consider a setting recently formulated as "open-world imitation from observation," where a robot imitates a manipulation task from a single video of human demonstration [4–6]. This process would facilitate users in effortlessly demonstrating a task to a robot and enable the humanoid robot to acquire new skills quickly.

Enabling humanoids to imitate from single videos presents a great challenge. A major challenge is that the videos do not have labels for robot actions. Prior works tackle this challenge by optimizing robot actions to reconstruct the future object motion trajectories [4, 5], but they are limited to single-arm tabletop manipulators. Therefore, optimization-based approach becomes computationally expensive for humanoids due to their high degrees of freedom and joint redundancy [7]. The similar kinematic structure shared by humans and humanoids allows for an alternative approach of retargeting, which directly translates human motions to humanoids [8, 9]. However, most retargeting techniques focus primarily on mimicking free-space body motions [10–14], lacking the awareness of object contexts for manipulation tasks. To address this shortcoming, we introduce the concept of "object-aware retargeting." By incorporating object awareness into the retargeting process, the resulting humanoid motions can be efficiently adapted to the locations of objects in open-ended environments.

We introduce OKAMI (**O**bject-aware **K**inematic ret**A**rgeting for humanoid **M**anipulation **I**mitation), an object-aware retargeting method enables a humanoid with two dexterous hands to imitate manipulation behaviors from a single RGB-D video demonstration. OKAMI is a two-stage process that retargets the human motions to the humanoid robot that accomplishes the task across varying initial conditions. The first stage processes the video to generate a reference manipulation plan for the subsequent stage, where the humanoid motion is synthesized through motion retargeting that adapts to the object locations during deployment.

OKAMI includes two key designs: The first design is an open-world vision pipeline that identifies task-relevant objects and reconstructs human motions from the video, and localizes task-relevant objects during evaluation. Localizing objects at test time also enables motion retargeting to adapt to different backgrounds or new instances of the same object categories, allowing the systematic generalization of the policy across varied visual conditions. The second design is the factorized process for retargeting, where we retarget the body motions and hand poses separately. We first retarget the body motions from the reference plan in the task space, and then warp the retargeted trajectory given the location of task-relevant objects. Then, the trajectory of body joints is obtained through inverse kinematics. Then, OKAMI directly maps the joint angles of fingers from the plan onto the dexterous hands, reproducing hand-object interaction. With object-aware retargeting, OKAMI policies are able to achieve systematic generalization across various spatial layouts of objects.

We evaluate OKAMI policies by providing video demonstrations of diverse tasks that cover various object interactions such as picking, placing, pushing, and pouring. We show that OKAMI policies achieve 71.66% task success rates averaged across all tasks in the experiment, outperforming the baseline by 58.33% on the selected two tasks. Qualitatively, we demonstrate that our humanoid robot is able to complete the demonstrated tasks in the real-world environments. In summary, our contributions of OKAMI are three-fold:

1. OKAMI enables a humanoid robot to mimic human behaviors from a single video to accomplish tasks. Its object-aware retargeting process generates feasible motions of the humanoid robot while adapting the motions to target object locations at test time.
2. OKAMI uses foundation models to identify task-relevant objects without additional human inputs. Their common-sense reasoning ability identifies task-relevant objects even if they are not directly in contact with other objects or the robot hands, therefore being able to imitate more diverse tasks than prior work.

3. We validate OKAMI's systematic generalization capabilities on humanoid robot hardware. OKAMI policies enable real-robot deployment in natural environments with different visual backgrounds, unseen object layouts, and new instances of task-relevant objects.

2 Related Work

Humanoid Robot Control. A large body of literature has studied controlling humanoid robots to complete locomotion or manipulation tasks [10, 12, 15]. Methods like motion planning or optimal control typically require a perfect physics model of the humanoid robot and are often computationally expensive [11, 12, 16]. People have explored using the sim-to-real paradigm, where they train reinforcement learning agents in simulation with domain randomization so that the policies can be transferred robustly. However, such a method is typically limited by the simulation tasks that can be created and often only limited to the locomotion tasks [10], whereas the simulation of manipulation tasks is hard to design, not to mention the reward functions. Using human data makes humanoid robot control easier, given the similar kinematic structures between humans and humanoids. The control can be done through teleoperation, using either motion capture suits [9, 12, 17–21], telexistence cockpits [22–26], VR devices [1, 27, 28], or using RGB video to track human motion [15]. However, such remote control requires real-time control of human teleoperators, posing both great mental and physical stress on the teleoperators. Instead, we focus on the setting where a robot watches the human perform a manipulation task in an RGB-D video. While existing literature has explored such an imitation setup in the scope of tabletop manipulation [4–6], we are the first to study the problem within the scope of humanoid manipulation.

Learning From Demonstrations / Imitation Learning. Imitation Learning has progressed significantly in learning vision-based robot manipulation with high sample efficiency [29–40]. Prior works have shown that with dozens of demonstrations, a robot can learn a visuomotor policy that completes various tasks, ranging from long-horizon manipulation tasks [30–32] to dexterous manipulation [33–35]. However, collecting demonstrations often requires expertise in using teleoperation devices, creating barriers to usability. Another line of work focuses on one-shot imitation learning [36, 37] or imitating from a single demonstration [38–40]. However, they either require additional data collection during a meta-training stage or still require teleoperation. Recently, people have shifted their focus towards imitating a single video demonstration without ground-truth label [4–6]. This problem was recently defined as "open-world imitation from observation" [4]. However, unlike prior works that explicitly abstract away the embodiment motions due to the kinematic differences between humans and robot arms, we exploit the embodiment information due to the embodiment similarity between human bodies and humanoid robots. In this work, OKAMI focuses on what we call object-aware retargeting. It is a method that adapts the motion of human bodies to humanoid robots so that we can achieve humanoid robot imitation.

Motion Retargeting. Motion retargeting has been long studied for adapting the motion of a person or a character to another character[8]. Retargeting has a wide application in computer graphics and 3D vision communities, where literature has extensively studied how to retarget human motions to human digital avatars [41–43]. This technique has been extended to robotics, where researchers focus on how to reuse the motions of a human and recreate similar behaviors on a humanoid robot or other robots with anthropomorphism. Rich liteartures have investigated how to do retargeting with a variety of methodology, such as optimization-based (QP, motion planning, IK) [11, 12, 17, 44], geometric-based (affine mapping, etc.) [45], and learning-based [10, 13, 15]. These methods have been successfully used in generating quadruped locomotion, loco-manipulation, humanoid locomotion, manipulation, and loco-manipulation. However, these retargeting methods have been used in teleoperation systems in the scope of manipulation tasks, as they lack a vision pipeline that allows the robot to adapt to object locations automatically. In this work, we connect the retargeting process with open-world vision, endowing the retargeting process with object awareness so that the robot mimics the human motions from a video demonstration and adapts to the object locations at test time.

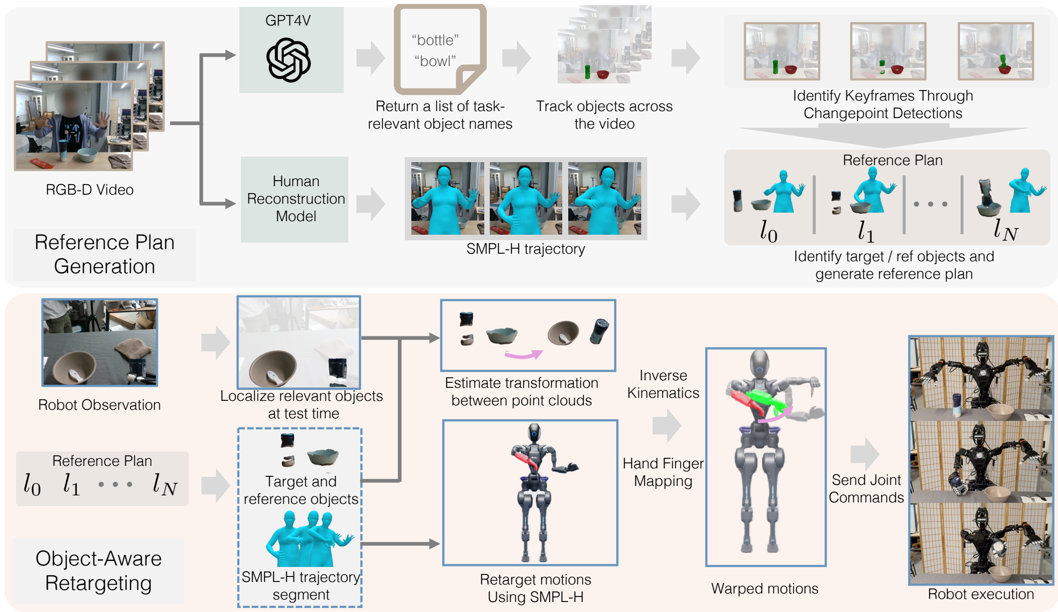

Figure 2: **Overview of OKAMI**. OKAMI is a two-staged method that enables a humanoid robot to imitate a manipulation task from a single human video. In the first stage, OKAMI generates a reference plan for subsequent manipulation, and the plan generation process uses GPT4V and multiple large vision models. In the second stage, OKAMI follows the reference plan, where it retargets human motions onto the humanoid with object awareness for each step of the plan. The retargeted motions are converted into robot joint configurations, and the humanoid robot follows the joint configurations to complete the demonstrated manipulation task.

3 OKAMI

In this work, we introduce OKAMI, a two-staged method that tackles "open-world imitation from observation." OKAMI first generates a *reference plan* using the object locations and reconstructed human motions from a given RGB-D video; then it retargets the human motions trajectories to the humanoid robot while adapting the trajectories based on new locations of the objects. Figure 2 illustrates the whole pipeline. We first formulate the problem of humanoid manipulation under "open-world imitation from observation." Then, following the formulation, we introduce the two stages of OKAMI: reference plan generation and object-aware retargeting.

3.1 Problem Formulation

We formulate a humanoid manipulation task as a discrete-time Markov Decision Process defined by a tuple: $M = (S, A, P, R, \gamma, \mu)$, where S is the state space, A is the action space, $P(\cdot|s, a)$ is the transition probability, $R(s)$ is the reward function, $\gamma \in [0, 1)$ is the discount factor, and μ is the initial state distribution. In our context, S is the space of raw RGB-D observations that capture both the robot and object states, A is the space of the motion commands for the humanoid robot, R is the sparse reward function that returns 1 when a task is complete. The objective of solving a task is to find a policy π that maximizes the expected task success rates from a wide range of initial configurations drawn from μ at test time.

In this paper, we consider the setting of "open-world imitation from observation" [4] in the scope of humanoid manipulation. In this setting, the robot system takes a recorded RGB-D human video, V as input, and returns a policy π that generates humanoid motion commands to complete the task as demonstrated in V. This setting is "open-world" as the robot does not have prior knowledge or ground-truth access to the categories or physical states of objects involved in the task, and it is "from observation" in the sense that video V does not come with any ground-truth robot actions. In this setting, a policy execution is considered successful if the state matches the states of the final frame from V. For all the tasks we evaluate, the success conditions are described in Appendix B.1. In this paper, two more assumptions are made about V: all the image frames in V capture the human bodies, and the camera view of shooting V is static throughout the recording.

3.2 Reference Plan Generation

To enable object-aware retargeting, OKAMI first needs to generate a reference plan for the humanoid robot to follow. Here, we describe how the reference plan is generated. To this end, OKAMI needs to understand what objects are involved and how humans move the objects in the demonstrated task, which are described first before we introduce the plan generation step.

Identify and Localize Task-Relevant Objects. Imitating a manipulation task requires the robot to understand what objects to interact with in order to complete the task. However, identifying task-relevant objects from pure images is a nontrivial challenge. While prior works use unsupervised approach to identify the objects [46, 47], they often assume simple visual backgrounds. Other alternatives require additional linguistic inputs from humans, inducing extra annotation cost from the user [48, 49]. Instead, we observe that most objects in everyday tasks are covered by common sense knowledge, where state-of-the-art Vision-Language Models (VLMs) such as GPT4V have internalized such knowledge through pre-training on internet data. Based on such observation, we leverage the power of GPT4V to identify task-relevant objects directly from the video demonstration V. Concretely, OKAMI samples the RGB image frames from V and prompting GPT4V with the concatenated image of the sampled frames (Appendix A.2 describe the details of test prompt we use to query the object names from GPT4V). GPT4V returns a list of texts that describe the names of task-relevant objects in V. Subsequently, OKAMI uses Grounded-SAM [50] with the list of object names to segments the objects on the first frame of V, and then track their locations across the entire video by propagating the first frame segmentation throughout the images using Cutie [51]. In the end, OKAMI localizes the task-relevant objects in V, which is the cornerstone for all subsequent steps.

Reconstruct Human Motions. As mentioned in Section 1, retargeting human motions to the humanoid has great potential to generate feasible actions for humanoids due to their human-like embodiments. However, the video demonstration V does not come with annotations on the human motions. To fill in the gap of missing data, we use a pre-trained vision model that can reconstruct 3D human models from in-the-wild videos (More details about training human reconstruction model are provided in Appendix A.1). The model outputs a sequence of SMPL-H (Skinned Multi-Person Linear Model with Hands) features [52], which capture the human body and hand poses throughout the video. From the trajectory of SMPL-H models, we obtain the estimated full-body poses, which include locations of body joints in the task space with respect to the human pelvis, and hand poses in joint configurations that describe how a hand interacts with an object. With the SMPL-H trajectories, OKAMI is able to retarget the human motions to the humanoids, which will be explained in Section 3.3. One advantage of using SMPL-H representation is that it captures human body poses while being invariant across humans with different demographics, and SMPL-H representation is easy to retarget motions to the humanoid robot that has different sizes from the human. As our experiments show, OKAMI is able to handle variations across different demonstrations.

Generate a Plan From V. From the previous two steps, the robot has the notion of both task-relevant objects and how human manipulate the objects. However, naively warping the entire human motion trajectory based on object locations doom to fail. Instead, OKAMI needs to identify the subgoals in V such that we can warp segment of trajectories conditioning on the location of the object that is associated with a subgoal.

We begin by performing temporal segmentations on the tracked object motions using changepoint detection, allowing us to identify subgoals. Next, we identify the target objects and reference objects for achieving each subgoal. This process is accomplished using a hybrid module that combines low-level point clouds to identify contacts and high-level common sense reasoning to understand objects that are not directly in contact (e.g., In a pouring task, the container is relevant to the task but never touched by the hand nor the cup).

Once the subgoals and associated objects are determined, we generate a reference plan, represented as $\{l_0, l_1, \dots, l_N\}$, where each step l_i corresponds to an identified keyframe. Each step stores a three-element tuple (o_{target}, $o_{reference}$, $\tau^{\text{SMPL}}_{t_i:t_{i+1}}$), which are the point clouds of the target object, the reference object and the SMPL-H trajectory segment between two keyframes, respectively. Note that $o_{reference}$ can be null if there is no spatial reference required for a step (e.g., grasping an object or closing a drawer as opposed the placing, where reference object is required). All the point clouds are obtained by back-projecting the segmented objects from RGB images using depth images [53].

3.3 Object-Aware Retargeting

Given a reference plan generated from the video demonstration, the humanoid robot follows the plan to imitate the demonstrated task in V. The robot follows each step l_i in the plan, where it first localizes the task-relevant objects, and retargets the corresponding segment of SMPL-H trajectory onto the humanoid while taking into account the target and reference objects. Then the retargeted trajectories are converted to the joint configuration trajectory using inverse kinematics for the robot hardware to execute. This process repeats until all the steps are executed and we evaluate if a rollout is successful or not following the success conditions of each task, as explained in Appendix B.1.

Localize Objects at Test Time. The reference plan is executed step by step, with each step containing a tuple of information about the target object, reference object, and the corresponding subgoal-bounded SMPL-H trajectory. To adapt the plan to the test-time environment, we localize the objects specified in the tuple using the robot's current observation. By extracting 3D point clouds of the objects from the robot's perception system, we can accurately track their positions and orientations. Localizing the objects at test-time paves the way for OKAMI to achieve systematic generalization across various visual conditions, including different backgrounds, and with new instances of task-relevant objects.

Retarget Human Motions to the Humanoid. The key aspect of object-awareness in our approach is the ability to adapt to new locations of objects. Once OKAMI localizes the objects in the observation, we develop a retargeting process that adapts humanoid motions to the object locations. Specifically, we employ a factorized process that separates the retargeting of the arm and hand motions. In this process, OKAMI first adapts the arm motions to the object locations so that the fingers of the hands are placed within the object-centric coordinate frame. Then OKAMI only needs to retarget fingers in the joint configuration to mimic how the human interacts with objects with their hands.

Concretely, the retargeting process begins by mapping the human body motions from the task space to the humanoid robot. This process involves scaling and adjusting the trajectories to account for the differences in size and proportion between the human and the robot. Next, OKAMI warps the retargeted trajectory based on the locations of the objects observed at test time. It essentially "bends" the trajectory to ensure that the robot's arm reaches the objects in their new positions while maintaining the overall trajectory shape of the demonstrated motions, making humanoid motions look natural. Specifically, there are two cases we consider for warping the trajectory: the first case is when there are no changes to the relational state between the target and the reference object or no reference object exists. In that case, we only warp the trajectory conditioning on the locations of the target object; the second case is where the relation state changes, meaning the trajectory needs to be conditioned on the reference object location.

Once the arm trajectory is warped, we use inverse kinematics to solve a sequence of joint configurations for the arms. At the same time, we retarget the human's hand poses to the robot in the configuration space. This means that we map the joint angles of the human hand to the corresponding joint angles of the robot's hand, ensuring that the robot can replicate the fine-grained manipulations demonstrated by the human. Together, we have the trajectory of full-body joint configurations for the real robot hardware to execute using a low-level robot controller.

Since the retargeting of arm motions between the human and the humanoid is affine, the retargeting process naturally allows us to scale and adjust motions given demonstrators with different demographics such as heights. By adapting the arm trajectories to the object locations and retargeting the

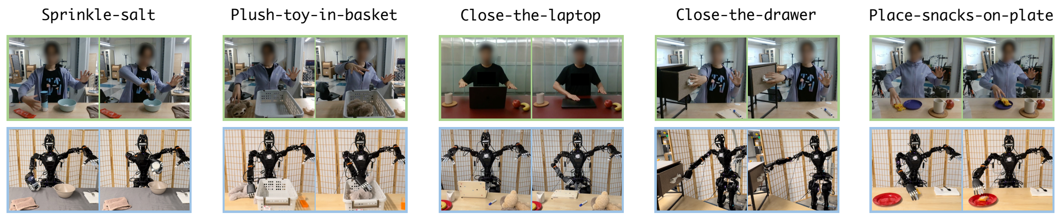

Figure 3: Visualization of initial and final frames of both human demonstrations and robot rollouts for all tasks.

hand poses independently, OKAMI's factorized process of retargeting achieves systematic generalization across various spatial layouts.

4 Experiments

Our experiments are designed to answer the following research question: (1) Is OKAMI effective for a humanoid robot to imitate diverse manipulation tasks by watching single videos of human demonstrations? (2) Is it critical in OKAMI to retarget the body motions of demonstrators to the humanoid robot instead of only retargeting based on object locations? (3) Does OKAMI keep the good performance consistently given videos of the same task demonstrated by users with diverse demographics?

4.1 Experimental Setup

Tasks. Here we describe the tasks we choose. 1) `Plush-toy-in-basket`: placing a plush toy in the basket; 2) `Sprinkle-salt`: sprinkling a bit of salt into the bowl; 3) `Close-the-drawer`: pushing the drawer in to close it; 4) `Close-the-laptop`: closing the lid of the laptop; 5) `Place-snacks-on-plate`: placing a bag of snacks on the plate. We select these five tasks that cover all kinds of manipulation behaviors: `Plush-toy-in-basket` and `Place-snacks-on-plate` require pick-and-place behaviors of daily objects; `Sprinkle-salt` is the task that covers pouring behavior; `Close-the-drawer` and `Close-the-laptop` require the humanoid to interact with articulated objects, which is a common interaction exist in daily environments.

Hardware Setup. We use Fourier-GR1 as the real robot hardware evaluation. The robot is equipped with two 6-DoF Inspire dexterous hands. For both video recording and robot camera observation, we use the D435i Intel RealSense camera. In all our experiments, we use a joint position controller that operates at 400Hz. To avoid jerky movements, we command the joint position targets at 40Hz and interpolate the commands to 400Hz trajectories.

Evaluation Protocol. We evaluate 12 trials for each task. The locations of the objects are initialized within the intersection of the robot camera's view and the humanoid arms' reachable range. The tasks are evaluated on a tabletop workspace with multiple objects, including both task-relevant objects and various other objects. Further, we test new object generalization on `Place-snacks-on-plate`, `Plush-toy-in-basket`, and `Sprinkle-salt` tasks, changing the involved plate, snack bag, plush toy, and bowl to other instances of the same type.

Baselines. We compare our result with a baseline ORION [4]. Since ORION was proposed for parallel-jaw gripper, we cannot directly apply it in our experiments. To evaluate ORION in the humanoid experiments, we've made minimal modifications: we estimate the palm trajectory using the SMPL-H trajectories, and warp the trajectory conditioning on the new object locations. The warped trajectory is used in the subsequent inverse kinematics for computing robot joint configurations.

4.2 Quantitative Results

To answer question (1), we evaluate the policies of our method across 5 different tasks (introduced in the experimental setup section), which cover diverse behaviors such as daily pick-place, pouring, and manipulation of articulated objects. The result is shown in Figure 4(a). In our experiment,

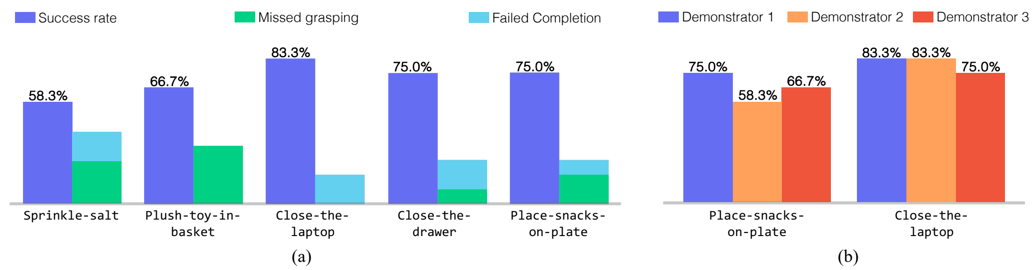

Figure 4: (a) Evaluation of OKAMI over all five tasks, including the success rates and the quantification of failed trials, separated by failure mode. (b) Evaluation of OKAMI using videos from different demonstrations. Demonstrator 1 is the main person recording videos for all evaluations in (a).

we randomly initialize the object locations, so that the robot needs to adapt to the locations of the objects. This result supports the design of OKAMI and shows its effectiveness in achieving systematic generalizations over different visual and spatial conditions.

To answer question (2), we compare OKAMI against ORION on two representative tasks, `Place-snacks-on-plate` and `Close-the-laptop`. OKAMI achieves 75.0% and 83.3% success rates, respectively, while ORION only achieves 0.0% and 41.2%, respectively. In the comparison experiment, OKAMI differs from ORION in that ORION does not condition on the human body poses. The outperforming result suggests the importance of retargeting the body motion of the human demonstrators onto the humanoid when imitating from human videos.

To answer question (3), we conduct a controlled experiment of recording videos of different demonstrators and test if OKAMI policies maintain good performance across different video inputs. Same as the previous experiment, we evaluate OKAMI on `Place-snacks-on-plate` and `Close-the-laptop` task. The result is shown in Figure 4(b). We show that for the task `Close-the-laptop`, there is no statistical significance in performance change. As for task `Place-snacks-on-plate`, while the evaluation maintains above 50%, the worst policy performance is 16.7% worse than the best policy performance. After looking into the video recording, we find that the motion of demonstrator 2 is relatively faster than the other two demonstrators, and faster motions create noisy estimation of motion when doing human model reconstruction. Overall, OKAMI is able to maintain reasonably good performance given videos from different demonstrators, but there is room for improvements to handle such variety.

5 Conclusion

This paper introduces OKAMI that enables a humanoid robot to imitate a single RGB-D human video demonstration. At the core of OKAMI is object-aware retargeting, which retargets the human motions onto the humanoid robot and adapts the motions to the object locations. OKAMI consists of two stages to realize object-aware retargeting. The first stage is generating a reference plan for manipulation from the video. The second stage is used for retargeting, where OKAMI retargets the arm motions in the task space and the finger motions in the joint configuration space. Our experiments validate the design of OKAMI, showing the systematic generalization of OKAMI policies.

Limitations. The focus of OKAMI is on the upper body motion retargeting of humanoid robots, particularly for manipulation tasks within tabletop workspaces. A promising future direction is to include lower body retargeting that enable locomotion behaviors during video imtiation. To enable full-body loco-manipulation, whole-body motion controller needs to be imnplemented as oppposed to the joint position controller we used in OKAMI.

Additionally, we focus on using RGB-D data in OKAMI, which prevents us from using in-the-wild internet videos recorded in RGB. Extending OKAMI will be another promising direction for future works.

References

[1] M. Seo, S. Han, K. Sim, S. H. Bang, C. Gonzalez, L. Sentis, and Y. Zhu. Deep imitation learning for humanoid loco-manipulation through human teleoperation. In *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, 2023.

[2] Y. Matsuura, K. Kawaharazuka, N. Hiraoka, K. Kojima, K. Okada, and M. Inaba. Development of a whole-body work imitation learning system by a biped and bi-armed humanoid. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10374–10381. IEEE, 2023.

[3] T. Asfour, P. Azad, F. Gyarfas, and R. Dillmann. Imitation learning of dual-arm manipulation tasks in humanoid robots. *International journal of humanoid robotics*, 5(02):183–202, 2008.

[4] 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.

[5] N. Heppert, M. Argus, T. Welschehold, T. Brox, and A. Valada. Ditto: Demonstration imitation by trajectory transformation. *arXiv preprint arXiv:2403.15203*, 2024.

[6] D. Guo. Learning multi-step manipulation tasks from a single human demonstration. *arXiv preprint arXiv:2312.15346*, 2023.

[7] T. Asfour and R. Dillmann. Human-like motion of a humanoid robot arm based on a closed-form solution of the inverse kinematics problem. In *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453)*, volume 2, pages 1407–1412. IEEE, 2003.

[8] M. Retargetting. Retargetting motion to new characters.

[9] K. Darvish, Y. Tirupachuri, G. Romualdi, L. Rapetti, D. Ferigo, F. J. A. Chavez, and D. Pucci. Whole-body geometric retargeting for humanoid robots. In *2019 IEEE-RAS 19th International Conference on Humanoid Robots (Humanoids)*, pages 679–686. IEEE, 2019.

[10] X. Cheng, Y. Ji, J. Chen, R. Yang, G. Yang, and X. Wang. Expressive whole-body control for humanoid robots. *arXiv preprint arXiv:2402.16796*, 2024.

[11] S. Nakaoka, A. Nakazawa, F. Kanehiro, K. Kaneko, M. Morisawa, and K. Ikeuchi. Task model of lower body motion for a biped humanoid robot to imitate human dances. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3157–3162. IEEE, 2005.

[12] K. Hu, C. Ott, and D. Lee. Online human walking imitation in task and joint space based on quadratic programming. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3458–3464. IEEE, 2014.

[13] S. Choi, M. K. Pan, and J. Kim. Nonparametric motion retargeting for humanoid robots on shared latent space. In *Robotics: Science and Systems*, 2020.

[14] E. Demircan, T. Besier, S. Menon, and O. Khatib. Human motion reconstruction and synthesis of human skills. In *Advances in Robot Kinematics: Motion in Man and Machine: Motion in Man and Machine*, pages 283–292. Springer, 2010.

[15] T. He, Z. Luo, W. Xiao, C. Zhang, K. Kitani, C. Liu, and G. Shi. Learning human-to-humanoid real-time whole-body teleoperation. In *arXiv*, 2024.

[16] A. Escande, N. Mansard, and P.-B. Wieber. Hierarchical quadratic programming: Fast online humanoid-robot motion generation. *The International Journal of Robotics Research*, 33(7): 1006–1028, 2014.

[17] L. Penco, N. Scianca, V. Modugno, L. Lanari, G. Oriolo, and S. Ivaldi. A multimode teleoperation framework for humanoid loco-manipulation: An application for the icub robot. *IEEE Robotics & Automation Magazine*, 26(4):73–82, 2019.

[18] D. Kim, B.-J. You, and S.-R. Oh. Whole body motion control framework for arbitrarily and simultaneously assigned upper-body tasks and walking motion. *Modeling, Simulation and Optimization of Bipedal Walking*, pages 87–98, 2013.

[19] A. Di Fava, K. Bouyarmane, K. Chappellet, E. Ruffaldi, and A. Kheddar. Multi-contact motion retargeting from human to humanoid robot. In *2016 IEEE-RAS 16th international conference on humanoid robots (humanoids)*, pages 1081–1086. IEEE, 2016.

[20] M. Arduengo, A. Arduengo, A. Colomé, J. Lobo-Prat, and C. Torras. Human to robot whole-body motion transfer. In *2020 IEEE-RAS 20th International Conference on Humanoid Robots (Humanoids)*, pages 299–305. IEEE, 2021.

[21] R. Cisneros, M. Benallegue, K. Kaneko, H. Kaminaga, G. Caron, A. Tanguy, R. Singh, L. Sun, A. Dallard, C. Fournier, et al. Team janus humanoid avatar: A cybernetic avatar to embody human telepresence. In *Toward Robot Avatars: Perspectives on the ANA Avatar XPRIZE Competition, RSS Workshop*, volume 3, 2022.

[22] S. Tachi, K. Komoriya, K. Sawada, T. Nishiyama, T. Itoko, M. Kobayashi, and K. Inoue. Telexistence cockpit for humanoid robot control. *Advanced Robotics*, 17(3):199–217, 2003.

[23] J. Ramos and S. Kim. Humanoid dynamic synchronization through whole-body bilateral feedback teleoperation. *IEEE Transactions on Robotics*, 34(4):953–965, 2018.

[24] Y. Ishiguro, T. Makabe, Y. Nagamatsu, Y. Kojio, K. Kojima, F. Sugai, Y. Kakiuchi, K. Okada, and M. Inaba. Bilateral humanoid teleoperation system using whole-body exoskeleton cockpit tablis. *IEEE Robotics and Automation Letters*, 5(4):6419–6426, 2020.

[25] F. Abi-Farrajl, B. Henze, A. Werner, M. Panzirsch, C. Ott, and M. A. Roa. Humanoid teleoperation using task-relevant haptic feedback. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5010–5017, 2018. doi:10.1109/IROS.2018.8593521.

[26] M. Schwarz, C. Lenz, A. Rochow, M. Schreiber, and S. Behnke. Nimbro avatar: Interactive immersive telepresence with force-feedback telemanipulation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5312–5319. IEEE, 2021.

[27] M. Hirschmanner, C. Tsiourti, T. Patten, and M. Vincze. Virtual reality teleoperation of a humanoid robot using markerless human upper body pose imitation. in 2019 ieee-ras 19th international conference on humanoid robots (humanoids), 2019.

[28] D. Lim, D. Kim, and J. Park. Online telemanipulation framework on humanoid for both manipulation and imitation. *2022 19th International Conference on Ubiquitous Robots (UR)*, pages 8–15, 2022. URL `https://api.semanticscholar.org/CorpusID:250577582`.

[29] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, and R. Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. *arXiv preprint arXiv:2108.03298*, 2021.

[30] A. Mandlekar, D. Xu, R. Martín-Martín, S. Savarese, and L. Fei-Fei. Learning to generalize across long-horizon tasks from human demonstrations. *arXiv preprint arXiv:2003.06085*, 2020.

[31] C. Wang, L. Fan, J. Sun, R. Zhang, L. Fei-Fei, D. Xu, Y. Zhu, and A. Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play. *arXiv preprint arXiv:2302.12422*, 2023.

[32] Y. Zhu, A. Joshi, P. Stone, and Y. Zhu. Viola: Imitation learning for vision-based manipulation with object proposal priors. *arXiv preprint arXiv:2210.11339*, 2022.

[33] C. Wang, H. Shi, W. Wang, R. Zhang, L. Fei-Fei, and C. K. Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. *arXiv preprint arXiv:2403.07788*, 2024.

[34] T. Lin, Y. Zhang, Q. Li, H. Qi, B. Yi, S. Levine, and J. Malik. Learning visuotactile skills with two multifingered hands. *arXiv preprint arXiv:2404.16823*, 2024.

[35] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu. 3d diffusion policy. *arXiv preprint arXiv:2403.03954*, 2024.

[36] M. Chang and S. Gupta. One-shot visual imitation via attributed waypoints and demonstration augmentation. *arXiv preprint arXiv:2302.04856*, 2023.

[37] T. Yu, P. Abbeel, S. Levine, and C. Finn. One-shot composition of vision-based skills from demonstration. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2643–2650. IEEE, 2019.

[38] E. Valassakis, G. Papagiannis, N. Di Palo, and E. Johns. Demonstrate once, imitate immediately (dome): Learning visual servoing for one-shot imitation learning. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8614–8621. IEEE, 2022.

[39] E. Johns. Coarse-to-fine imitation learning: Robot manipulation from a single demonstration. In *2021 IEEE international conference on robotics and automation (ICRA)*, pages 4613–4619. IEEE, 2021.

[40] N. Di Palo and E. Johns. Learning multi-stage tasks with one demonstration via self-replay. In *Conference on Robot Learning*, pages 1180–1189. PMLR, 2022.

[41] Z. Luo, J. Cao, K. Kitani, W. Xu, et al. Perpetual humanoid control for real-time simulated avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10895–10904, 2023.

[42] X. B. Peng, Z. Ma, P. Abbeel, S. Levine, and A. Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (ToG)*, 40(4): 1–20, 2021.

[43] B. Jiang, X. Chen, W. Liu, J. Yu, G. Yu, and T. Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36, 2024.

[44] S. Kuindersma, R. Deits, M. Fallon, A. Valenzuela, H. Dai, F. Permenter, T. Koolen, P. Marion, and R. Tedrake. Optimization-based locomotion planning, estimation, and control design for the atlas humanoid robot. *Autonomous robots*, 40:429–455, 2016.

[45] Y. Liang, W. Li, Y. Wang, R. Xiong, Y. Mao, and J. Zhang. Dynamic movement primitive based motion retargeting for dual-arm sign language motions. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8195–8201. IEEE, 2021.

[46] S. Caelles, J. Pont-Tuset, F. Perazzi, A. Montes, K.-K. Maninis, and L. Van Gool. The 2019 davis challenge on vos: Unsupervised multi-object segmentation. *arXiv preprint arXiv:1905.00737*, 2019.

[47] Y. Huang, J. Yuan, C. Kim, P. Pradhan, B. Chen, L. Fuxin, and T. Hermans. Out of sight, still in mind: Reasoning and planning about unobserved objects with video tracking enabled memory models. *arXiv preprint arXiv:2309.15278*, 2023.

[48] Y. Zhu, Z. Jiang, P. Stone, and Y. Zhu. Learning generalizable manipulation policies with object-centric 3d representations. In *7th Annual Conference on Robot Learning*, 2023.

[49] 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.

[50] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[51] H. K. Cheng, S. W. Oh, B. Price, J.-Y. Lee, and A. Schwing. Putting the object back into video object segmentation. In *arXiv*, 2023.

[52] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019.

[53] Q.-Y. Zhou, J. Park, and V. Koltun. Open3d: A modern library for 3d data processing. *arXiv preprint arXiv:1801.09847*, 2018.

[54] S. Goel, G. Pavlakos, J. Rajasegaran, A. Kanazawa, and J. Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023.

[55] Y. Xu, J. Zhang, Q. Zhang, and D. Tao. ViTPose++: Vision transformer for generic body pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[56] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024.

[57] V. Ye, G. Pavlakos, J. Malik, and A. Kanazawa. Decoupling human and camera motion from videos in the wild. In *CVPR*, 2023.

[58] J. Romero, D. Tzionas, and M. J. Black. Embodied hands. *ACM Transactions on Graphics*, 36 (6):1–17, 2017.