Self-Supervised Multisensory Pretraining for Contact-Rich Robot Reinforcement Learning

Rickmer Krohn Vignesh Prasad Gabriele Tiboni Prof. Georgia Chalvatzaki
Interactive Robot Perception & Learning
Department of Computer Science
TU Darmstadt, Germany
{rickmer.krohn, vignesh.prasad,
gabriele.tiboni, georgia.chalvatzaki}@tu-darmstadt.de

Abstract

An important element of learning contact-rich manipulation for robots is leveraging the synergy between heterogeneous sensor modalities such as vision, force, and proprioception while adapting to sensory perturbations and dynamic changes. In such multisensory settings, Reinforcement Learning (RL) faces caveats arising from varying sensory feature distributions and their changing importance depending on the task phase. To this end, taking a leaf out of Multimodal representation learning, pretraining yields itself as a natural way to learn robust cross-modal feature representations useful for different downstream tasks. In this work, we propose MultiSensory Dynamic Pretraining (MSDP), a novel framework for learning multisensory representations tailored for task-oriented policy learning via masked autoencoding coupled with self-supervised forward dynamics objectives to shape features from multiple different sensors. Using our pretraining approach, we demonstrate how using a simple cross-attention between a learnable task-specific embedding and frozen multisensory embeddings yields consistently strong performance on downstream tasks. Our approach simplifies the handling of sensory interactions, allowing the agent to focus entirely on mastering the task. Our method demonstrates accelerated learning and robust performance under diverse perturbations, including sensor noise, and changes in object dynamics. Evaluations across multiple challenging contact-rich robot manipulation tasks showcase the effectiveness and robustness of MSDP. Our framework's modular pretraining process supports various sensor combinations, providing a simple and robust solution for complex manipulation tasks.

1 Introduction

Reinforcement Learning (RL) has shown impressive successes in learning complex tasks from games like Atari [43], to locomotion [64], vision-based manipulation [63], and multi-sensory peg-insertion [30, 31, 52]. However, incorporating multiple sensor modalities, especially in more complex contact-rich robotic manipulation tasks, remains a challenge for RL, due to the heterogeneous dynamics of different sensor modalities. Additionally, the importance of each input modality changes during the execution of a manipulation task, e.g., coarse scene understanding from visual input to fine-grained force feedback when in contact. Thus, robotic agents need to learn how to dynamically focus on the most relevant sensory information while adapting to perturbations and dynamic changes in the environment. This capability, of effective *sensor fusion*, is a long-studied problem in the field of robotics for various tasks ranging from control [16, 26, 59, 6], manipulation [25, 58], localization, and navigation [2], but remains underexplored in RL settings.

18th European Workshop on Reinforcement Learning (EWRL 2025).

To this end, Imitation Learning approaches [35, 23, 50] have shown promise in utilizing multisensory data for learning skills, but require experts collecting informative data. The challenge of data collection becomes particularly pronounced in tasks involving objects with varying or uncertain properties, such as mass or friction. In such cases, substantial and often expensive data acquisition efforts are required to develop a robust policy capable of generalizing effectively across diverse contextual variations [5]. Instead, RL offers a promising avenue to mitigate the challenges associated with extensive data collection. Self-directed exploration in RL facilitates the development of adaptable strategies that can be generalized across diverse object properties and contexts. These advantages position RL as a compelling approach for learning multisensory contact-rich manipulation [30, 31, 8, 38]. Recent advances in multimodal learning [3, 18] have shown the advantages of reconstructing masked or noised inputs to learn robust cross-modal representations for downstream tasks. Such masking-based self-supervision approaches have been shown to improve the network's robustness [53, 37, 27]. Masking inputs, whether at the sensory or embedding level, facilitates the learning of robust representations when trained on extensive datasets [11, 18, 4]. Recent works have also explored how such masked multisensory pretraining, can enhance representation learning for contact-rich manipulation [52, 38].

In this paper, we propose ***M**ulti**S**ensory* ***D**ynamic* ***P**retraining* (MSDP), a novel framework for learning multisensory representations that combine masked autoencoding with dynamics-aware self-supervised predictions to capture the interactions between different sensory modalities. Our intuitive use of masking sensor embeddings, resulting in cross-sensor predictions, enables learning a robust and synergistic fusion of sensory data in a way that is robust to sensory noise and dropout while enabling seamless integration of multiple modalities.

We decouple the representation learning and downstream RL and demonstrate how our pretraining approach yields meaningful sensor representations that can enable effective learning via a simple yet effective architecture wherein we use cross-attention between a learnable embedding and the frozen multisensory embeddings to obtain a rich task-specific representation for the critic. The policy acts directly on the multisensory representation to effectively solve downstream robotic tasks. Our experimental results demonstrate that the proposed approach yields effective representations for accelerating RL on a variety of contact-rich manipulation tasks in a manner robust to sensory noise and changing object dynamics. By focusing on global representation extraction and leveraging multisensory masked autoencoding, MSDP advances the capabilities to use RL in challenging contact-rich tasks, where dynamics matter.

To summarize, the key contributions of our work are twofold: (i) we develop an effective pretraining strategy that merges masked autoencoding and forward dynamics modeling and (ii) introduce a novel multisensory architecture that allows task specific feature extraction to achieve effective sensory fusion to form an expressive multisensory latent for robotic Reinforcement Learning of contact-rich manipulation tasks. Our findings pave the way for more adaptive and resilient RL-agents to handle multiple input modalities to master complex manipulation tasks.

2 Related Work

2.1 Reinforcement Learning for Contact-rich Manipulation

Tasks where a robotic manipulator has to interact with its environment, either directly or indirectly via a tool, require a good understanding of the interaction forces that shape the task at hand. Solving contact-rich tasks often relies on an accurate estimate of the force and dynamics of the task at hand. When such estimates are available, classical control approaches can be adapted to solve the task [56, 57]. However, such approaches often require a lot of hand-crafting to tune the controller parameters for various tasks. While adapting force/impedance controller parameters can be learned from demonstration [47], to adapt to more unstructured environments, and to better handle unseen or difficult contact dynamics, Reinforcement Learning presents itself as an ideal candidate that can learn via interactions with the environment [45]. In the context of contact-rich manipulation, various works have employed RL to learn contact-based policies on a variety of tasks [13]. Additionally, the advent of visuomotor policy learning [33] opened up new horizons for abstracting out accurate state estimation needed for such tasks. A more in-depth review on reinforcement learning for contact-rich robotic manipulation tasks can be found in [13, 54, 39].

2.2 Multimodal Self-Supervision for Robotics

Lee et. al. [30, 31] learn a multimodal representation from Vision, Force-Torque, and Proprioception via MLP-fusion and multiple Self-Supervised objectives. The frozen representation leads to a robust representation for RL to solve multiple Peg Insertion tasks. [8] extended the Vision Transformer [12] with a Force-Torque sensor to solve a variety of contact-rich tasks. They additionally incorporate the self-supervised objectives from [31] to shape a representation using SLAC [29]. Mejia et. al. [42] pretrain an audio-encoder [44] to combine Vision and Audio via a Transformer Decoder for Imitation Learning. The audio signal from the contact microphone provides rich feedback for various manipulation tasks. To also account for different sensor frequencies [50] developed a multi-resolution policy based on pretrained Vision Language Models to increase inference time using Proprioception and Force-Torque.

2.3 Deep Sensor Fusion

A robust representation is essential to integrate and process various sensory inputs to ensure stable and efficient learning in multisensory RL. Previous architectures often focused on straightforward latent fusion approaches by concatenating the various representations for downstream tasks [32, 36, 34, 19, 22]. Feng et. al. [14] take this a step further by adding a subgoal-aware weighting for learning the stage-wise importance of difference sensors. Alternatively, contrastive learning approaches [10, 40] emphasize feature alignment, rather than fusion, to learn a shared latent representation between multiple modalities. However, distilling task-relevant features while maintaining sensor-specific information can be challenging. In contrast, inspired by the success of masked token prediction [3], recent works have also explored how masked multisensory pre-training can enhance representation learning for contact-rich manipulation [52, 38].

3 Preliminaries

We define a multisensorial POMDP as a tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{O}_{MS},\ \mathcal{A},\ \mathcal{R},\ \mathcal{P},\ \gamma \rangle$ [28], where $\mathcal{S}$ is the true state of the environment. The agent does not have access to $\mathcal{S}$ and observes it through multiple sensor observations, which we specify as Proprioception $\mathcal{O}_P \in \mathbb{R}^{14}$, Force Torque (FT) $\mathcal{O}_{FT} \in \mathbb{R}^{4\times 6}$ and Vision $O_V \in \mathbb{R}^{64\times 64\times 3}$. The sensors build up the multisensory Observation-space $\mathcal{O}_{MS} = [\mathcal{O}_P,\ \mathcal{O}_{FT},\ \mathcal{O}_V]$. $\mathcal{A}$ is the action space, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the reward function, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is the transition kernel, and $\gamma \in [0, 1)$ is the discount factor. Our goal is to learn a policy $\pi : \mathcal{O}_{MS} \to \mathcal{A}$ that maximizes the discounted sum of rewards $\mathbb{E}_\pi[\sum_t^\infty \gamma^t r(s_t, a_t, s_{t+1})]$.

We specify Proprioception as the joint position and velocities, the Force Torque sensor feedbacks 4 wrench readings per POMDP-step and the vision observation is obtained through an external RGB-Camera. We aim to extract an expressive, fused representation from all sensors in order to solve contact-rich manipulation tasks with RL. We only provide the current multisensory observation and no history. We use the off-policy algorithm SAC [20, 21], however we stress that our approach is compatible with any actor-critic RL algorithm.

4 Multisensory Dynamic Pretraining

We present **M**ulti**S**ensory **D**ynamic **P**retraining (MSDP), a novel framework for representation learning that builds upon Masked Autoencoders [24] and transformer-based architectures, tailored to enhance contact-rich robot reinforcement learning tasks that require perception through multiple sensor modalities. MSDP introduces a modular architecture to seamlessly handle varying input sensors, and a multisensory masking scheme to promote rich cross-modal representations, i.e., to retain knowledge about the task and the environment dynamics even in the absence or perturbation of one or more modalities. Particularly, we modify the original masked autoencoding objective to predict the next observation instead of reconstructing the current one, resulting in dynamics-aware features. Furthermore, we shed light on the different ways to map pre-trained transformer embeddings to input states for downstream RL tasks, referred to as *latent bridging*, a problem often overlooked in practice by the community.

In this context, we propose employing a simple cross-attention layer to obtain expressive task-specific features from frozen multisensory embeddings for the critic representation. The actor on the other

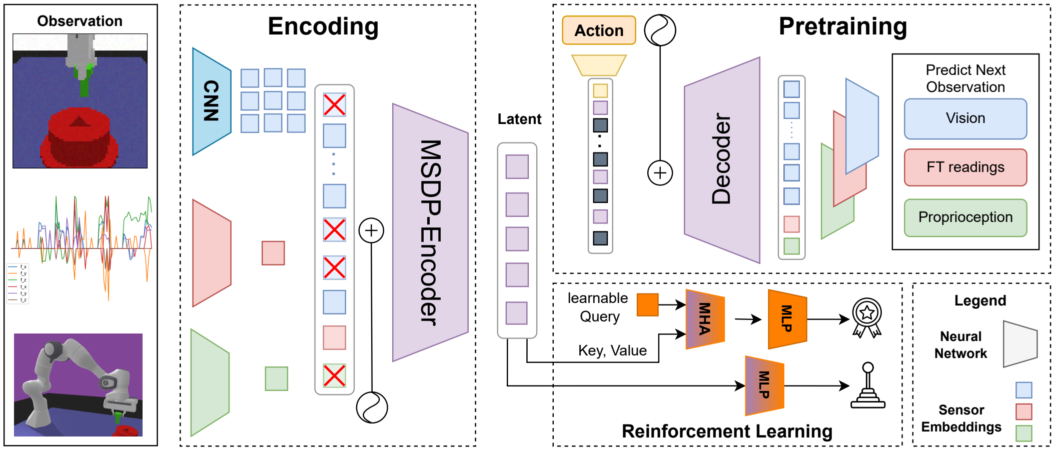


Figure 1: The MSDP framework with MSDP-Encoder (left), Pretraining (top right) and down stream RL (botton-right): The current multisensory observation gets projected with a CNN-stem (vision) and linear layers (Force Torque and Proprioception) to the embedding space. The MSDP-encoder fuses all sensor embeddings to form our expressive multisensory latent representation. The encoder is trained via the decoder and next sensor observation prediction from a subset of sensor embeddings. This pretraining results in dynamic cross sensor prediction, shaping and fusing sensor representations. For downstream RL we extract multisensory task-specific features via a single cross-attention layer for the critic and via pooling for the actor. Sensor embeddings are only masked during pretraining. Our Framework provides dynamic and expressive multisensory features for contact-rich manipulation tasks.

hand receives a special pooling of the sensor embeddings to ensure a stable representation. Both mechanisms offers a fixed low-dimensional observation for the downstream RL-agent.

Overall, our framework decouples representation learning and downstream RL to ensure rich feature extraction and modular sensor modalities, while offering stability and compact downstream representations. Particularly, notice that our architecture is scalable w.r.t. the number of input modalities and does not make any assumptions on the learning pipeline of the downstream RL task.

As a result, our pretraining phase on limited offline data leads to rich representations that may be directly used to solve complex contact-rich RL tasks from high-dimensional input (including images) in less then $500k$ environment steps.

We structure the description of our method as follows: Section 4.1 describes the neural network architecture; Section 4.2 explains the multisensory pretraining scheme for representation learning; Section 4.3 describes the details of latent bridging and policy learning for downstream robotic tasks. Figure 1 displays an overview of the proposed framework.

4.1 MSDP Architecture

Given the inputs from multiple sensor modalities, such as vision, proprioception and end-effector forces, we encode each modality using a separate layer/network referred to as *sensor encoders*.

Particularly for the vision input, we use a CNN-stem similar to [51], which allows us to mask at embedding- rather than pixel-level. Masked object positions can not be retrieved in pixel-space. The CNN-stem also introduces redundancy as vision-embeddings have overlapping receptive fields and, most importantly, stabilizes training [61]. Furthermore, compared to a patchify-stem, it reduces the burden of the multisensory encoder to extract vision features, while fusing with other modalities. We use a linear projection to encode force-torque readings and proprioception given their low dimensionalities.

Subsequently, we incorporate position encoding for the embedded sensor features (or *tokens*) to contextualize the position and modality of each token. Rather than using fixed positional/modality encoding [55, 7] we use learnable embeddings to encode position and modality together, adding only 1.2% of overall parameters. Once we obtain embedding representations from all sensors, we randomly mask out a subset of those sensor embeddings and feed the remaining to our multisensory transformer encoder, resulting in our multisensory latent embeddings. The self attention mechanism [55] of the encoder leads to dynamic multisensory fusion promoted by the representation objective described in Section 4.2.

A learnable mask embedding is added to the randomly masked out embedding positions, next to the multisensory embeddings from the encoder to generate the original number of embeddings. Additionally, an action embedding from a linear projection is added to make the decoder action-conditioned. We reapply our positional/modality encoding before feeding all embeddings to the decoder. In the decoder embeddings exchange information, especially for the mask embedding, before being feed to separate decoder heads to fulfill the Representation Learning objective described in Section 4.2. We use a shared linear projection for all vision embeddings to reconstruct their corresponding patch. Examples of Vision prediction can be found in Figure 12. The extraction of a global multisensory representation is explained in Section 4.3.

4.2 Representation Learning

Our representation learning objective is based on multimodal masked autoencoding [3, 18]. The objective is to predict the next sensor observation from a subset of sensor embeddings. Vision as a global sensing modality has a high number of embeddings and needs to extract information about other, potentially masked, sensors e.g. identifying contact to estimate force. Other sensors are beneficial to predict the next (vision) observation as e.g. Proprioception indicates the robot position. This representation objective results in cross-sensor prediction thus leading to fusion of all modalities. Furthermore, the decoder is conditioned on action and needs to extract dynamic action-related features important for downstream RL. We denote our method **MSDP-P** (Prediction) wherein we reconstruct the next observation $||\mathbf{O}_{t+1} - \Phi(\mathbf{O}_t, \mathbf{A}_t)||^2$ and **MSDP-R** (Reconstruction) where we reconstruct the current observation $||\mathbf{O}_t - \Phi(\mathbf{O}_t)||^2$, where $\Phi(.)$ denotes the network prediction conditioned on the current observation $\mathbf{O}_t$ and optionally action $\mathbf{A}_t$. $|| \cdot ||^2$ denotes the Mean Squared Error.

Non-transformer based architectures (see baselines in Sec. 5.1) trained on reconstruction introduce a bottleneck limiting expressiveness as task irrelevant features (e.g. background) are encoded. Our embedding-based latent rather yields redundant features promoting robustness rather than a bottleneck limiting performance.

4.3 Policy Learning

From the pretrained encoder, we receive expressive multisensory embeddings given the available sensors. Extracting a global representation from those embeddings for downstream task solving is a crucial and often underdiscussed design choice, while having a considerable impact on performance. To address this, we depart from works that naively extract the "CLS"-embedding from the high-dimensional embeddings, analogous to the Vision Transformer [12, 60, 48, 41] and, instead, propose a asymmetric *latent bridging* strategy between actor and critic to obtain a compact representation. The critic uses a single cross-attention layer with a learnable query and the multisensory embeddings from the MSDP encoder as key's and value's (see Figure 1). It offers dynamic task-specific feature extraction (e.g. object positions, robot state, contact) over the task solving process. The fine-grained understanding of the environment leads to faster convergence compared to a global representation. The policy on the other hand does not profit from a cross-attention layer as it's destabilizing training. Instead, we mean pool all embeddings originating from the vision sensor before pooling all sensor embeddings, to account for the uneven number of embeddings. Pooling sensor embeddings results in a stable and parameter free latent bridging similiar to [51, 52]. This asymmetric representation between actor and critic, follows [17], where the actor benefits from a stable representation over task stages and the critic from a value specific representation. The cross-attention layer is trained solely by the critic resulting in task-specific feature extraction. We train with the off-policy algorithm SAC, where we incorporate the offline data for pre-training in our replay buffer. Our approach is working with any actor-critic RL algorithm.

5 Experiments and Results

In this Section, we present our experimental setting, consisting of two competitive baselines, training details and multisensory environments. Furthermore, we ablate the impact of sensor settings and latent bridging mechanisms. Finally, we investigate how pretraining with multiple sensors can enrich the vision representation for Downstream RL.

5.1 Baselines and Training Details

We compare our methods against two competitive baselines. Both Baselines extract sensor-specific features with a CNN for Vision and an MLP for Proprioception and Force-Torque readings. The **Concat** model fuses the concatenated features with a 2-layer MLP to form the multisensory latent representation [30, 31]. The **PoE** model generates separate means and variances for each sensor and fuses them with a Product of Expert approach. As the models are non transformer based we neglect masking and pretrain the encoders with reconstruction to form a multisensory latent. Both models use an identical decoder architecture for the representation objectives, where proprioception and force-torque signals are reconstructed from the latent representation using separate two-layer MLPs. Vision is reconstructed via a Deconvolutional-CNN. More details can be found in the appendix A.5.

Training Details: We collect 30,000 random samples from the environment and train each model with its respective representation learning objective (see Section 4.2) for 30,000 update steps. After pretraining, we freeze the encoder and train the on downstream task for 100 epochs. Each epoch consists of 5,000 interaction steps, totaling 500,000 RL update steps.

5.2 Multisensory Environments

Our multisensory environments are based on panda-gym [15] and Pybullet [9]. The tasks—Peg Insertion, Push Cube and Close Drawer Gently—are inspired by Minitouch [8, 49] and consist of 3 challenging manipulation tasks. Each task comes with multisensory observations consisting of proprioception (joint position and velocities), vision from an external RGB camera and a force-torque sensor located on the wrist of the robot. The action-space consists of the 3-dimensional end-effector displacement if not stated otherwise. Task details are provided below, with an overview of the environments available in Appendix A.1.

Peg Insertion: Insert the triangular peg (fixed at the endeffector) into the hole. The dense reward function is guiding the robot to place the peg above the hole before inserting it. Next to the position, the z-orientation of the triangular-peg needs to match the hole, leading to a 4-dimensional action space and complex insertion dynamics. We randomize the hole position and orientation. Furthermore, we add gaussian noise to the vision observation to mimic sensor noise and encourage the usage of the FT-sensor to solve the task.

Push Cube: Push an orange cube to a target location with a round peg held by the robot. Goal threshold is 3cm, while we provide a dense reward function. Next to the position of the cube, we randomize its mass and center of mass leading to changing object dynamics [46].

Close Drawer Gently: Instead of fully closing a drawer, the Task is successful, when the drawer is nearly closed, while having a small velocity. It requires fine-grained control to be solved. A standard close-drawer task formulation often results in forceful closing behavior, which is undesirable in real-world applications. We randomize the position and orientation of the drawer and the friction of its prismatic joint.

Figure 2 shows the performance of our MSDP-P and MSDP-R models compared to the proposed Baselines. We obtain superior performance in Peg Insertion and Push Cube indicating a rich representation for RL. Especially in Peg Insertion, our sensor-fusion provides fine-grained features to solve the task around 80% of the time after just 40 epochs. Our Baselines struggle to provide an appropriate representation for the given Task. As Push Cube is more Vision-focused and does not introduce vision noise, the PoE baseline can solve the task in a reasonable timeframe. Our method performs less prominent in Close Drawer Gently Task mainly due to simulation bottlenecks. We observed sudden drawer changes during contact limiting the ability of the agent to close the drawer smoothly in order to solve the task consistently. Regardless, our pretraining allows the agent to understand the complex sensor interactions faster and obtains a higher final performance / success rate

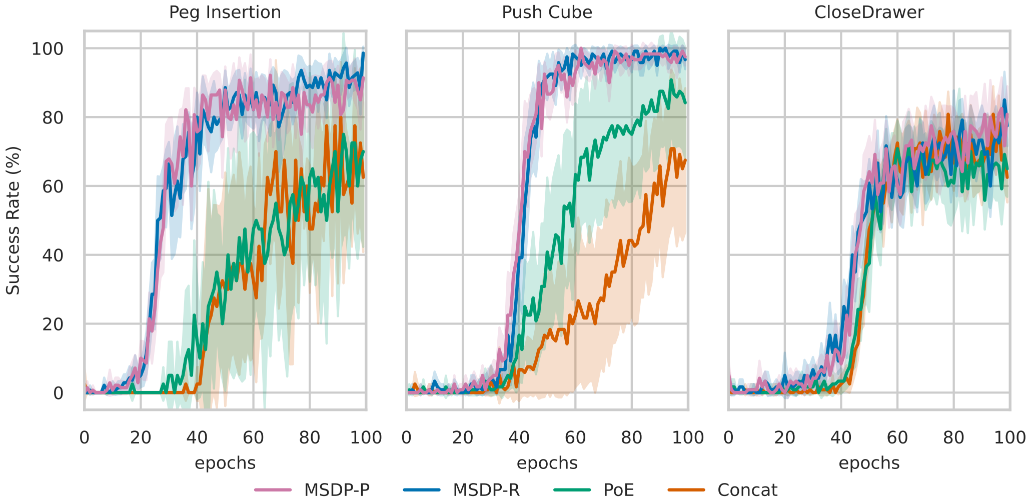


Figure 2: Performance comparison between MSDP-P, MSDP-R and the baselines in Peg Insertion, Push Cube and Close Drawer Gently. Our Method achieves the highest success rate across all Tasks.

(~79%) compared to the Baselines (~63%). Additionally, our models report higher force interactions (see Figure 11), indicating fast reaching of the drawer and movement adjustment once in contact. In the following sections we discuss sensor- and latent-ablations as well how multiple sensors can enrich the vision representation for the MSDP-P model. More results for MSDP-R are reported in Section A.3.

5.3 Sensor Ablation

To identify the contribution of each sensor we ablate them for the Peg Insertion and Push Cube environment for our MSDP-P model. As seen in Figure 3 the usage of all three sensor modalities obtain best performance. This is more prominent in the Peg Insertion Task, where Proprioception and the FT-sensor are crucial to insert the peg with limited vision features. The Push Cube task can be solved with Vision only, as identifying the cube position is crucial. The additional sensors still provide useful information especially, where the center of mass is shifted significantly.

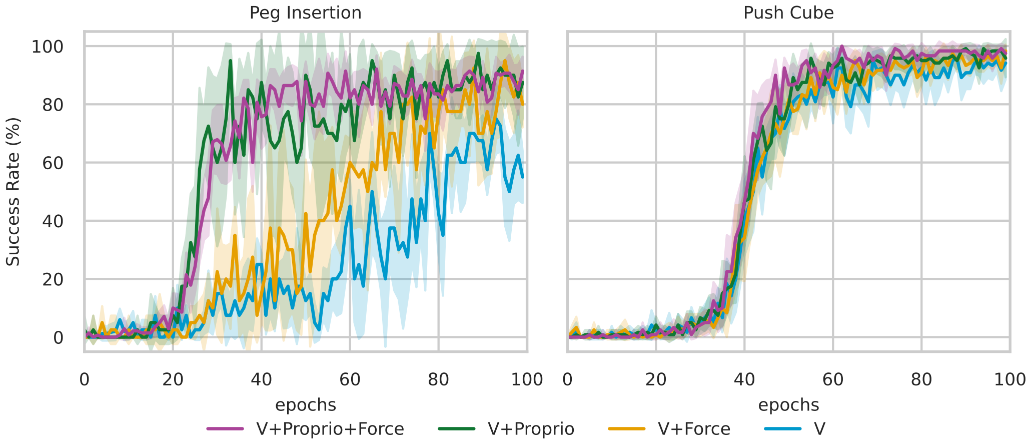


Figure 3: Sensor ablation of MSDP-P for Peg Insertion and Push Cube. The usage of all sensors leads to a rich representation indicated by the highest performance. This is especially prominent in Peg Insertion, where all sensors are needed to insert the peg consistently. "V" stand for Vision

Next to the performance gains, we observed that models in Peg Insertion without having access to the FT-modality asserts higher forces on the Hole (see Section A.4), indicating unwanted exploration actions, which may damage the robot or environment in the real world.

5.4 Latent bridging results

The latent bridging mechanism, described in Section 4.3, to obtain a compact representation from multisensory embeddings can have a big impact on performance. We compare our critic cross-attention extraction with common approaches. *CLS* is using the CLS-embedding of the encoder, commonly used in the Vision Transformer [12] or Imitation Learning [50]. In *Pooling* we mean-pool all sensor embeddings [51, 52]. To account for the different number of embeddings we first take the mean of all vision embeddings. To avoid possible dilution *Cat* provides the Concatenation of sensor embeddings [62], where we pool the vision embeddings before. The common latent bridging mechanisms do not introduce new learnable parameters, while our Cross attention layer only adds minimal parameters to extract task-related features.

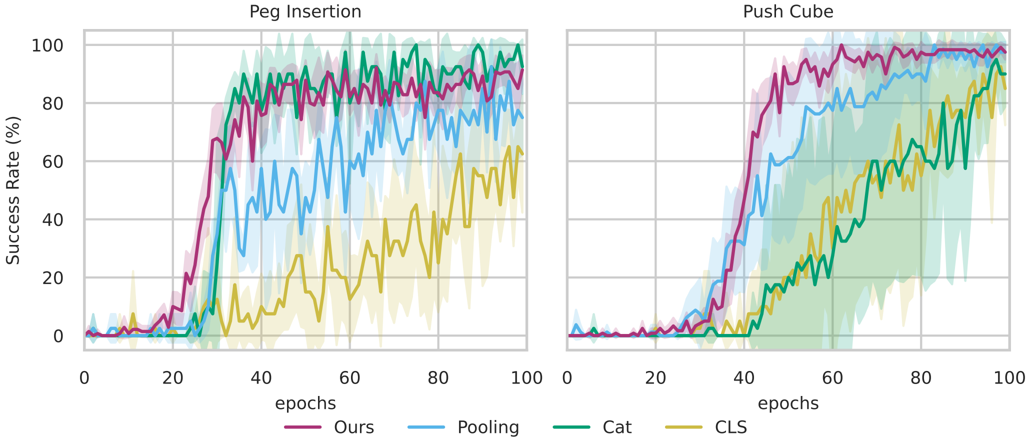


Figure 4: Peg Insertion and PushCube Latent bridging ablation

Figure 4 shows different latent bridging mechanisms on our MSDP-P model in Peg Insertion and Push Cube. Our critic cross-attention (CCA) successfully extracts a rich representation to learn the task in a fast manner. We highlight that all embeddings from the pre-trained encoder are multisensorial, even if we can assign their origin from a specific sensor. In Push Cube our mechanism can extract more fine-grained vision features resulting in superior performance. Figure 6 shows the attention map of CCA's latent bridging, focusing on the cube and its target. This confirms how our cross-attention extracts task specific features from multisensory embeddings. The Peg Insertion task demands detailed information about robot position and contact, which is also given by the *Cat* latent bridging. The *CLS* embedding doesn't contain rich features for downstream task learning, while *Pooling* takes longer to solve the task consistently as the agent needs to extract environment details from a potentially diluted representation.

5.5 Enriched Representation for Vision-based RL

In this Section, we investigate how low-dimensional sensor modalities can enrich the vision representation. Our MSDP transformer encoder allows for varying input length, offering us the option to pretrain the representation on more sensors and only use a subset, e.g. only Vision, for downstream task learning. We pretrain our MSDP model with different sensor configurations and only provide the vision modality during task solving. The latent is constructed by mean-pooling all vision embeddings.

Figure 5 shows how different pretraining settings can enhance vision-based RL in the Peg Insertion and Push Cube tasks. In all configurations, we pretrain using the MSDP-P objective. The vision representation in Peg Insertion is significantly improved by incorporating other sensors, increasing the success rate from 40% to over 70%. In contrast, the representation for the vision-focused Push Cube task degenerates when all sensor modalities are used during pretraining but not during dowmstream

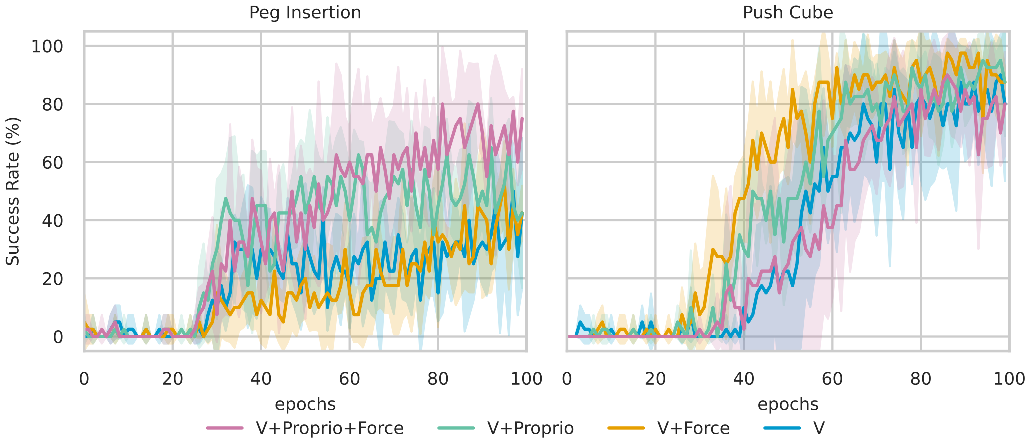


Figure 5: Peg Insertion and Push Cube, Comparisons of different Pretraining, Only Vision during RL. "V" stand for Vision

RL. During task execution, the observed vision distribution shifts from the random collected data during pretraining. The data shift in combination with the absence of other sensors, leads to corrupted features. However, using only one additional sensor (Proprioception or FT-sensor) during pretraining improves performance. Notably, the FT sensor adds beneficial features to the vision representation, as it's indicating cube contact.

6 Conclusion and Outlook

In this work, we propose MultiSensory Dynamic Pretraining (MSDP), a novel pre-training framework for learning multisensory representations for contact-rich manipulation tasks using masked autoencoding and self-supervised forward dynamics objectives. Specifically, MSDP learns to reconstruct sensory information of future timesteps in an action-conditioned manner from a subset of input sensor embeddings. MSDP captures the interplay between the different sensor modalities to learn a rich multisensory representation, which in combination with task specific feature extraction through cross attention leads to a superior performance in challenging contact-rich manipulation tasks. MSDP's architecture supports various sensor combinations and can shift each sensor's importance with its attention mechanism, allowing for task-oriented multisensory feature extraction. In the future, we plan to evaluate MSDP's performance in more challenging real-world settings, especially the robustness of the latent representation against sensor dropout or perturbation like noise. Moreover, the flexible and decoupled nature of our representation learning provides a meaningful direction on how to exploit heterogeneous multisensory data, which we aim to explore in a multi-task fashion as part of our future work.

References

[1] Samira Abnar and Willem Zuidema. Quantifying Attention Flow in Transformers, May 2020. arXiv:2005.00928 [cs].

[2] Mary B Alatise and Gerhard P Hancke. A review on challenges of autonomous mobile robot and sensor fusion methods. *IEEE Access*, 8:39830–39846, 2020.

[3] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. MultiMAE: Multi-modal Multi-task Masked Autoencoders, April 2022.

[4] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022.

[5] Suneel Belkhale, Yuchen Cui, and Dorsa Sadigh. Data quality in imitation learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[6] Isela Bonilla, Marco Mendoza, Emilio J Gonzalez-Galvan, Cesar Chavez-Olivares, Ambrocio Loredo-Flores, and Fernando Reyes. Path-tracking maneuvers with industrial robot manipulators using uncalibrated vision and impedance control. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(6):1716–1729, 2012.

[7] Xin Chen, Sam Toyer, Cody Wild, Scott Emmons, Ian Fischer, Kuang-Huei Lee, Neel Alex, Steven H. Wang, Ping Luo, Stuart Russell, Pieter Abbeel, and Rohin Shah. An Empirical Investigation of Representation Learning for Imitation, May 2022. arXiv:2205.07886.

[8] Yizhou Chen, Andrea Sipos, Mark Van der Merwe, and Nima Fazeli. Visuo-Tactile Transformers for Manipulation, September 2022. arXiv:2210.00121 [cs].

[9] Erwin Coumans and Yunfei Bai. PyBullet, a Python module for physics simulation for games, robotics and machine learning, 2016.

[10] Vedant Dave, Fotios Lygerakis, and Elmar Rueckert. Multimodal Visual-Tactile Representation Learning through Self-Supervised Contrastive Pre-Training, January 2024. arXiv:2401.12024 [cs].

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *CoRR*, abs/2010.11929, 2020. _eprint: 2010.11929.

[13] Íñigo Elguea-Aguinaco, Antonio Serrano-Muñoz, Dimitrios Chrysostomou, Ibai Inziarte-Hidalgo, Simon Bøgh, and Nestor Arana-Arexolaleiba. A review on reinforcement learning for contact-rich robotic manipulation tasks. *Robotics and Computer-Integrated Manufacturing*, 81:102517, 2023.

[14] Ruoxuan Feng, Di Hu, Wenke Ma, and Xuelong Li. Play to the score: Stage-guided dynamic multi-sensory fusion for robotic manipulation. In *8th Annual Conference on Robot Learning*, 2024.

[15] Quentin Gallouédec, Nicolas Cazin, Emmanuel Dellandréa, and Liming Chen. panda-gym: Open-Source Goal-Conditioned Environments for Robotic Learning. *4th Robot Learning Workshop: Self-Supervised and Lifelong Learning at NeurIPS*, 2021.

[16] Javier Gámez García, Anders Robertsson, Juan Gómez Ortega, and Rolf Johansson. Sensor fusion for compliant robot motion control. *IEEE Transactions on Robotics*, 24(2):430–441, 2008.

[17] Samuel Garcin, Trevor McInroe, Pablo Samuel Castro, Prakash Panangaden, Christopher G. Lucas, David Abel, and Stefano V. Albrecht. Studying the Interplay Between the Actor and Critic Representations in Reinforcement Learning, March 2025. arXiv:2503.06343 [cs].

[18] Xinyang Geng, Hao Liu, Lisa Lee, Dale Schuurmans, Sergey Levine, and Pieter Abbeel. Multimodal Masked Autoencoders Learn Transferable Representations, May 2022.

[19] Irmak Guzey, Ben Evans, Soumith Chintala, and Lerrel Pinto. Dexterity from touch: Self-supervised pre-training of tactile representations with robotic play. *arXiv preprint arXiv:2303.12076*, 2023.

[20] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor, 2018. _eprint: 1801.01290.

[21] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. Soft Actor-Critic Algorithms and Applications, 2019. _eprint: 1812.05905.

[22] Yunhai Han, Kelin Yu, Rahul Batra, Nathan Boyd, Chaitanya Mehta, Tuo Zhao, Yu She, Seth Hutchinson, and Ye Zhao. Learning generalizable vision-tactile robotic grasping strategy for deformable objects via transformer. *IEEE/ASME Transactions on Mechatronics*, 2024.

[23] Yilun Hao, Ruinan Wang, Zhangjie Cao, Zihan Wang, Yuchen Cui, and Dorsa Sadigh. Masked Imitation Learning: Discovering Environment-Invariant Modalities in Multimodal Demonstrations. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–7, October 2023. ISSN: 2153-0866.

[24] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners, December 2021. arXiv:2111.06377 [cs].

[25] Yingbai Hu, Zhijun Li, Guanglin Li, Peijiang Yuan, Chenguang Yang, and Rong Song. Development of sensory-motor fusion-based manipulation and grasping control for a robotic hand-eye system. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 47(7):1169–1180, 2016.

[26] Fouad F Khalil and Pierre Payeur. Dexterous robotic manipulation of deformable objects with multi-sensory feedback-a review. *Robot Manipulators Trends and Development*, (March 2010), 2010.

[27] Michael Kleinman, Alessandro Achille, and Stefano Soatto. Critical Learning Periods for Multisensory Integration in Deep Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24296–24305, June 2023.

[28] Hanna Kurniawati. Partially Observable Markov Decision Processes (POMDPs) and Robotics, July 2021. arXiv:2107.07599 [cs].

[29] Alex X. Lee, Anusha Nagabandi, Pieter Abbeel, and Sergey Levine. Stochastic Latent Actor-Critic: Deep Reinforcement Learning with a Latent Variable Model, October 2020. arXiv:1907.00953 [cs].

[30] Michelle A. Lee, Yuke Zhu, Krishnan Srinivasan, Parth Shah, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making Sense of Vision and Touch: Self-Supervised Learning of Multimodal Representations for Contact-Rich Tasks. *CoRR*, abs/1810.10191, 2018. _eprint: 1810.10191.

[31] Michelle A. Lee, Yuke Zhu, Peter Zachares, Matthew Tan, Krishnan Srinivasan, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making Sense of Vision and Touch: Learning Multimodal Representations for Contact-Rich Tasks, July 2019. arXiv:1907.13098 [cs].

[32] Michelle A Lee, Yuke Zhu, Peter Zachares, Matthew Tan, Krishnan Srinivasan, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Learning multimodal representations for contact-rich tasks. *IEEE Transactions on Robotics*, 36(3):582–596, 2020.

[33] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016.

[34] Ajian Li, Ruikai Liu, Xiansheng Yang, and Yunjiang Lou. Reinforcement learning strategy based on multimodal representations for high-precision assembly tasks. In *Intelligent Robotics and Applications: 14th International Conference, ICIRA 2021, Yantai, China, October 22–25, 2021, Proceedings, Part I 14*, pages 56–66. Springer, 2021.

[35] Hao Li, Yizhi Zhang, Junzhe Zhu, Shaoxiong Wang, Michelle A. Lee, Huazhe Xu, Edward Adelson, Li Fei-Fei, Ruohan Gao, and Jiajun Wu. See, Hear, and Feel: Smart Sensory Fusion for Robotic Manipulation, December 2022. arXiv:2212.03858 [cs].

[36] Hao Li, Yizhi Zhang, Junzhe Zhu, Shaoxiong Wang, Michelle A Lee, Huazhe Xu, Edward Adelson, Li Fei-Fei, Ruohan Gao, and Jiajun Wu. See, hear, and feel: Smart sensory fusion for robotic manipulation. In *Conference on Robot Learning*, pages 1368–1378. PMLR, 2023.

[37] Guan-Horng Liu, Avinash Siravuru, Sai Prabhakar, Manuela Veloso, and George Kantor. Learning End-to-end Multimodal Sensor Policies for Autonomous Navigation. 2017.

[38] Qingtao Liu, Zhengnan Sun, Yu Cui, Li Gaofeng, Qi Ye, and Jiming Chen. *Masked Visual-Tactile Pre-training for Robot Manipulation*. February 2024.

[39] Rongrong Liu, Florent Nageotte, Philippe Zanne, Michel de Mathelin, and Birgitta Dresp-Langley. Deep reinforcement learning for the control of robotic manipulation: a focussed mini-review. *Robotics*, 10(1):22, 2021.

[40] Fotios Lygerakis, Vedant Dave, and Elmar Rueckert. M2CURL: Sample-Efficient Multimodal Reinforcement Learning via Self-Supervised Representation Learning for Robotic Manipulation, June 2024. arXiv:2401.17032 [cs].

[41] Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Yecheng Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Pieter Abbeel, Jitendra Malik, Dhruv Batra, Yixin Lin, Oleksandr Maksymets, Aravind Rajeswaran, and Franziska Meier. Where are we in the search for an Artificial Visual Cortex for Embodied Intelligence?, February 2024. arXiv:2303.18240 [cs].

[42] Jared Mejia, Victoria Dean, Tess Hellebrekers, and Abhinav Gupta. Hearing Touch: Audio-Visual Pretraining for Contact-Rich Manipulation, May 2024. arXiv:2405.08576 [cs].

[43] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing Atari with Deep Reinforcement Learning, December 2013. arXiv:1312.5602 [cs].

[44] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-Visual Instance Discrimination with Cross-Modal Agreement, March 2021. arXiv:2004.12943 [cs].

[45] Mamix Nuttin and H Van Brussel. Learning the peg-into-hole assembly operation with a connectionist reinforcement technique. *Computers in Industry*, 33(1):101–109, 1997.

[46] Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-Real Transfer of Robotic Control with Dynamics Randomization. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3803–3810, May 2018. arXiv:1710.06537 [cs].

[47] Mattia Racca, Joni Pajarinen, Alberto Montebelli, and Ville Kyrki. Learning in-contact control strategies from demonstration. In *2016 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 688–695. IEEE, 2016.

[48] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-World Robot Learning with Masked Visual Pre-training, October 2022. arXiv:2210.03109 [cs].

[49] Sai Rajeswar, Cyril Ibrahim, Nitin Surya, Florian Golemo, David Vazquez, Aaron Courville, and Pedro O. Pinheiro. Haptics-based Curiosity for Sparse-reward Tasks. In Aleksandra Faust, David Hsu, and Gerhard Neumann, editors, *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 395–405. PMLR, November 2022.

[50] Saumya Saxena, Mohit Sharma, and Oliver Kroemer. MResT: Multi-Resolution Sensing for Real-Time Control with Vision-Language Models, January 2024. arXiv:2401.14502 [cs].

[51] Younggyo Seo, Danijar Hafner, Hao Liu, Fangchen Liu, Stephen James, Kimin Lee, and Pieter Abbeel. Masked World Models for Visual Control, May 2023. arXiv:2206.14244 [cs].

[52] Carmelo Sferrazza, Younggyo Seo, Hao Liu, Youngwoon Lee, and Pieter Abbeel. The Power of the Senses: Generalizable Manipulation from Vision and Touch through Masked Multimodal Learning, November 2023. arXiv:2311.00924 [cs].

[53] Skand Skand, Bikram Pandit, Chanho Kim, Li Fuxin, and Stefan Lee. Simple Masked Training Strategies Yield Control Policies That Are Robust to Sensor Failure. September 2024.

[54] Markku Suomalainen, Yiannis Karayiannidis, and Ville Kyrki. A survey of robot manipulation in contact. *Robotics and Autonomous Systems*, 156:104224, 2022.

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[56] Daniel E Whitney. Historical perspective and state of the art in robot force control. *The International Journal of Robotics Research*, 6(1):3–14, 1987.

[57] Daniel E Whitney et al. Quasi-static assembly of compliantly supported rigid parts. *Journal of Dynamic Systems, Measurement, and Control*, 104(1):65–77, 1982.

[58] Ziwei Xia, Zhen Deng, Bin Fang, Yiyong Yang, and Fuchun Sun. A review on sensory perception for dexterous robotic manipulation. *International Journal of Advanced Robotic Systems*, 19(2):17298806221095974, 2022.

[59] Di Xiao, Bijoy K Ghosh, Ning Xi, and Tzyh Jong Tarn. Sensor-based hybrid position/force control of a robot manipulator in an uncalibrated environment. *IEEE Transactions on control systems technology*, 8(4):635–645, 2000.

[60] Tete Xiao, Ilija Radosavovic, Trevor Darrell, and Jitendra Malik. Masked Visual Pre-training for Motor Control, March 2022. arXiv:2203.06173.

[61] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early Convolutions Help Transformers See Better, October 2021. arXiv:2106.14881 [cs].

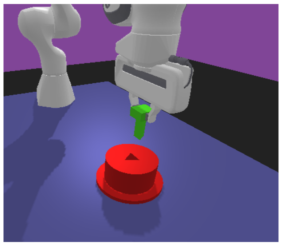
[62] Ruihan Yang, Minghao Zhang, Nicklas Hansen, Huazhe Xu, and Xiaolong Wang. Learning Vision-Guided Quadrupedal Locomotion End-to-End with Cross-Modal Transformers, May 2022. arXiv:2107.03996.

[63] Fangyi Zhang, Jürgen Leitner, Michael Milford, Ben Upcroft, and Peter Corke. Towards Vision-Based Deep Reinforcement Learning for Robotic Motion Control, November 2015. arXiv:1511.03791 [cs].

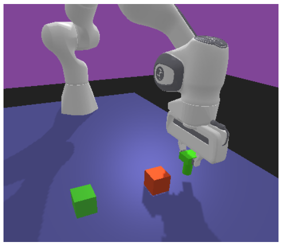
[64] Ziwen Zhuang, Zipeng Fu, Jianren Wang, Christopher Atkeson, Soeren Schwertfeger, Chelsea Finn, and Hang Zhao. Robot Parkour Learning, September 2023. arXiv:2309.05665 [cs].

A Technical Appendix

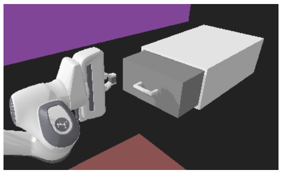
A.1 Environment Details



Peg Insertion: Insert the triangular peg (fixed at the end-effector) into the hole. The dense reward function got two stages. First: lead the peg to an area above the hole based on the xy-distance between peg and hole. Second, with a small orientation penalty: encourage the peg to lower its position inside the hole. Exact position is not indicated by the reward function. A high reward is achieved if the peg is fully inserted. Next to the position, the z-orientation of the triangular-peg is part of the 4-dimensional action space. We randomize the hole position in a 15cm box and vary the z-orientation in a span of 80 degrees.

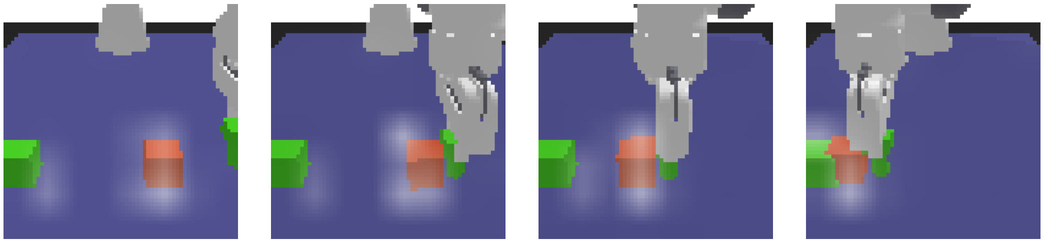


Push Cube: Push an orange cube to a target location with a round peg held by the robot. Goal threshold is 3cm, while we provide a dense reward function based on the distance between endeffector-cube and cube-goal. A high reward is achieved for task success. Next to the position of the cube, we randomize its mass and center of mass leading to changing object dynamics.



Close Drawer Gently: Instead of fully closing a drawer, the Task is successful, when the drawer is nearly closed, while having a small velocity. It requires fine-grained control to be solved. A standard close-drawer task formulation often results in forceful closing behavior, which is undesirable in real-world applications. We randomize in a 10cm x10cm x10cm box and orientation in a span of 30 degreees. Furthermore we vary the friction of the drawer's prismatic joint. The exact goal is to push the drawer between 3cm and 1cm openess, while having a velocity under 1cm/s. The dense reward is guiding the robot to the drawer and indicates the distance between desired and actual drawer state. Right drawer position get's rewarded as well as Task success.

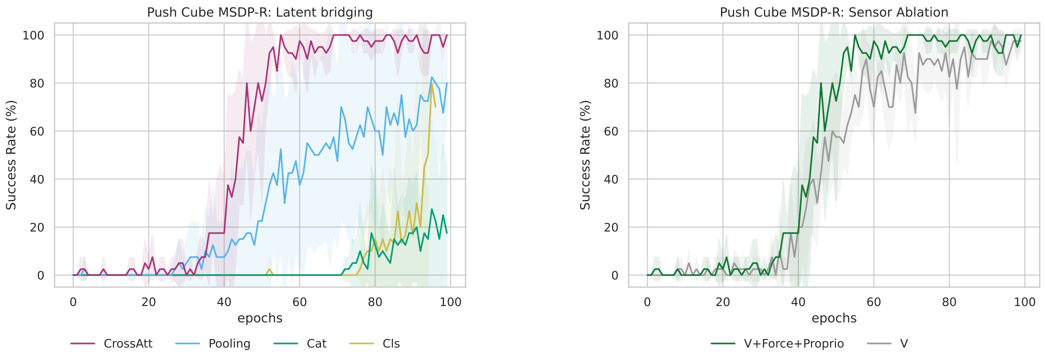
A.2 Cross-attention of critic in the Push Cube task



(a) timestep: 0 (b) timestep: 4 (c) timestep: 8 (d) timestep: 12

Figure 6: Critic ross-attention (*CCA*) maps in the Push Cube Task pretrained with MSDP-P. The critic is focusing on the cube and its target, indicating task relevant feature extraction.

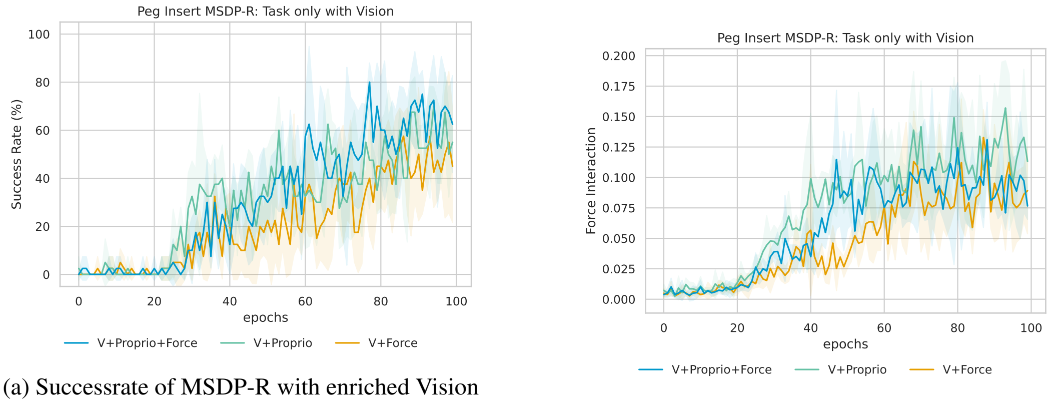
A.3 MSDP-R Results



(a) Latent ablation on Push Cube

(b) Sensor ablation on Push Cube

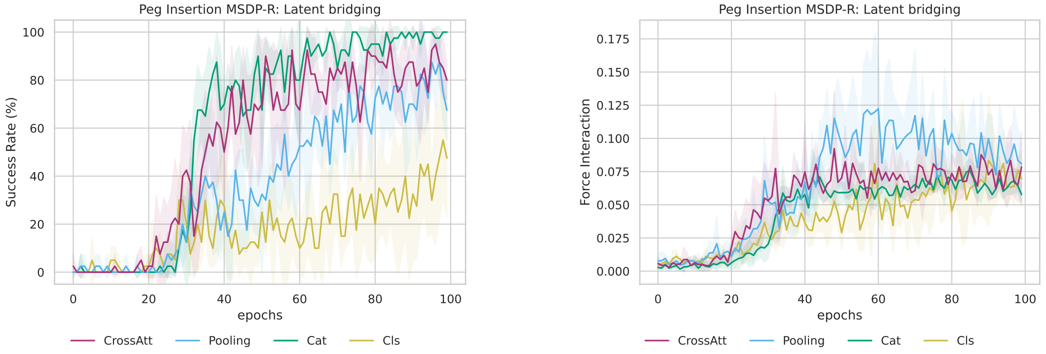
Figure 7: Latent and Sensor ablation for MSDP-R in Push Cube Task



(a) Successrate of MSDP-R with enriched Vision Representation

(b) Touch Oeginsert

Figure 8: Force Interaction of MSDP-R with enriched Vision Representation

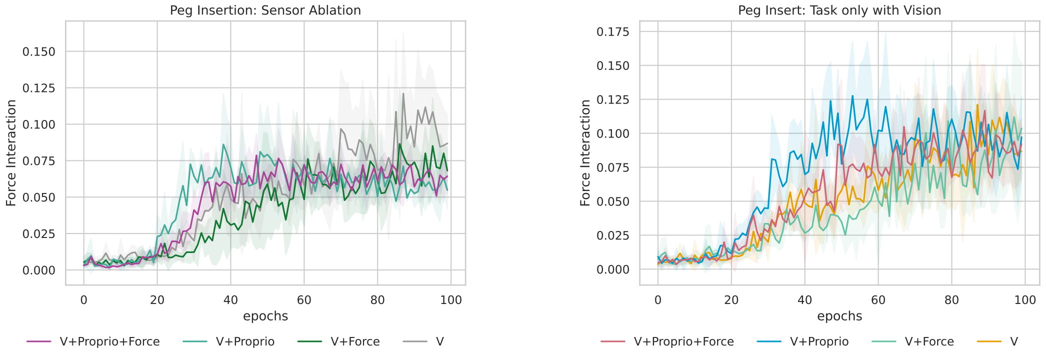


(a) Success rate

(b) Force Interaction

Figure 9: Latent Ablation of MSDP-R in Peg Insertion Task

A.4 Force Interactions



(a) Sensor ablation

(b) Rl only with vision

Figure 10: We define the force interaction as the standard deviation of the force-torque distribution over an episode. Agents with access to Vision and Prorioception during training (a) or pretraining (b) exihbit higher forces on the environment. Especially during exploration the missing force information is leading to sub-optimal actions, which may be dangerous in the real world.

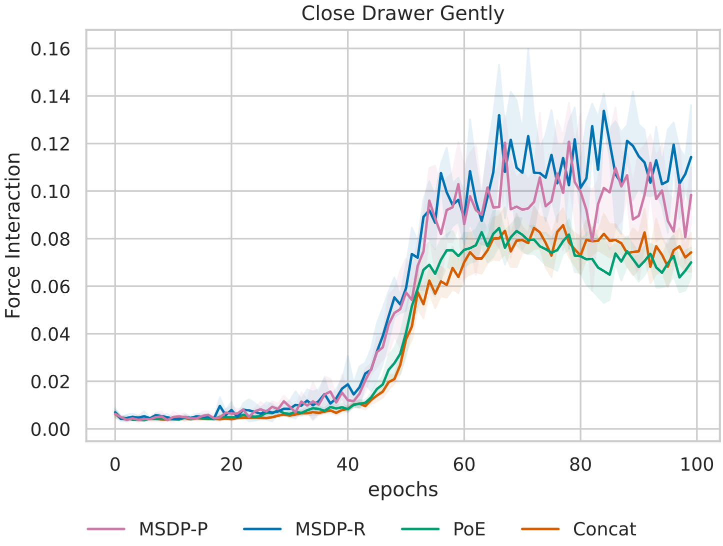


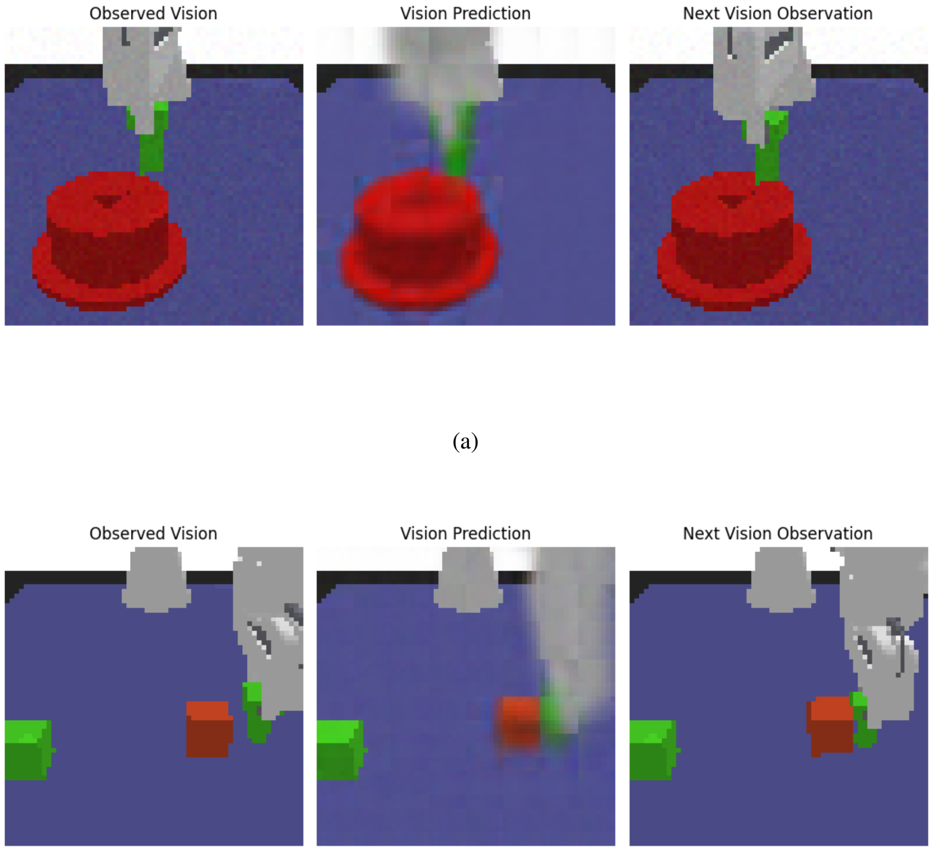
Figure 11: Force Interaction in the Close Drawer Gently Task. High Force Interaction indicates fast reaching of the drawer, while adjusting the movement once in contact. Our Methods archive comparable results with the baselines.

A.5 Baseline details

Following equation describes the latent generation from the PoE-baselines. From each sensor modality i we extract a normal distribution, which gets fused over each feature j.

$$\sigma_j^2 = \left(\sum_{i=1}^{3} \sigma_{ij}^2 \right)^{-1}, \quad \mu_j = \left(\sum_{i=1}^{3} \frac{\mu_{ij}}{\sigma_{ij}^2} \right) \left(\sum_{i=1}^{3} \sigma_{ij}^2 \right)^{-1}$$

A.6 Vision Predictions

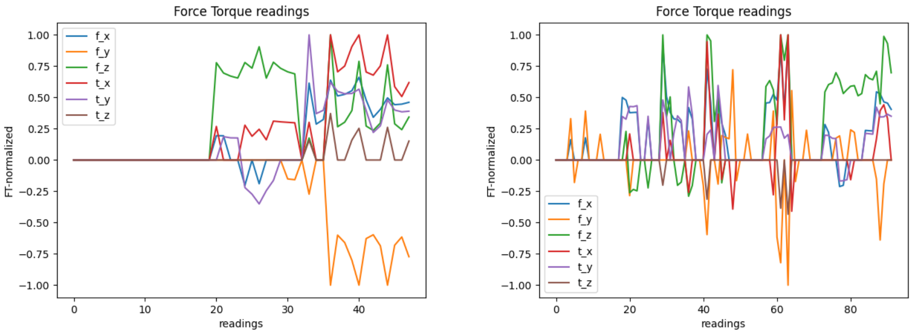


(a)

(b)

Figure 12: Vision observation prediction in (a) PegInsertion and (b) PushCube environment. Left: Observed Vision from our Multsensory Masked Autoencoder. Middle: The Vision Prediction. Right: The actual next Vision observation. Our model predicts correct peg movement under 60% masking ratio, while the shared linear projection layer introduces artifacts and is not able to predict fine-grained details.

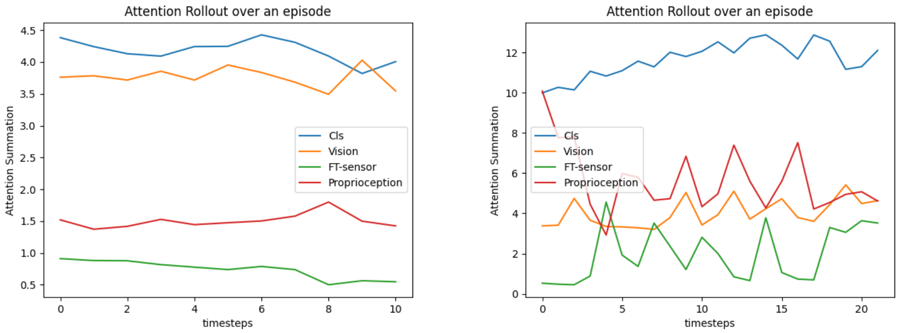
A.7 Force Observation



(a) (b)

Figure 13: Force Observation for (a) Peg Insertion and (b) Push Cube.

A.8 Attention MAE encoder



(a) (b)

Figure 14: Attention Rollout [1] for (a) Peg Insertion and (b) Push Cube. We report only the maximum attention on a single vision embedding to account for its high embedding number. The attention mechanism reported here is solely trained on the representation objective and is not an indicator on the sensor importance for the agent.