The Power of the Senses: Generalizable Manipulation from Vision and Touch through Masked Multimodal Learning

Carmelo Sferrazza[1], Younggyo Seo[2], Hao Liu[1], Youngwoon Lee[1], and Pieter Abbeel[1]

Abstract— **Humans rely on the synergy of their senses for most essential tasks. For tasks requiring object manipulation, we seamlessly and effectively exploit the complementarity of our senses of vision and touch. This paper draws inspiration from such capabilities and aims to find a systematic approach to fuse visual and tactile information in a reinforcement learning setting. We propose Masked Multimodal Learning (M3L), which jointly learns a policy and visual-tactile representations based on masked autoencoding. The representations jointly learned from vision and touch improve sample efficiency, and unlock generalization capabilities beyond those achievable through each of the senses separately. Remarkably, representations learned in a multimodal setting also benefit vision-only policies at test time. We evaluate M3L on three simulated environments with both visual and tactile observations: robotic insertion, door opening, and dexterous in-hand manipulation, demonstrating the benefits of learning a multimodal policy. Videos of the experiments are available at `https://sferrazza.cc/m3l_site`. Code will be released upon acceptance.**

I. Introduction

Humans are capable of exploiting the synergies and complementarities of their senses [1], [2], [3]. For example, when grasping an object, we fully rely on our sense of vision at first, since no physical feedback is available until contact is made. Once the object has been reached, visual feedback becomes partly or fully occluded by the human hand. Thus, vision-based reasoning is naturally replaced by rich touch feedback. Human reasoning and decision-making present uncountable similar examples, where different sensory modalities seamlessly cooperate with each other.

However, in robotic manipulation, vision and touch have mostly been studied independently, mainly due to the delayed development of tactile sensors compared to the widespread availability of high-performance visual sensing. While vision-based manipulation research has shown impressive achievements through modern machine learning approaches [4], [5], incorporating contact feedback with vision is crucial to broaden the capabilities of robotic manipulation, e.g., dealing with visual occlusion, manipulating fragile objects, and improving accuracy. Yet, a large part of touch-based manipulation research has so far focused on showcasing the potential of new high-resolution tactile sensors [6], [7], often limited to proof-of-concepts based on the assumption that visual sensing is unavailable.

In this paper, we propose Masked Multimodal Learning (M3L), which leverages both visual and tactile sensing modalities by systematically fusing them for manipulation tasks. Specifically, we focus on sample efficiency and generalization of reinforcement learning (RL) via multimodal representations extracted across vision and touch. To acquire such generalizable multimodal representations, we use a multimodal masked autoencoder (MAE) [8] that learns to extract condensed representations by optimizing a reconstruction loss based on raw visual and tactile observations, while simultaneously optimizing a policy that is conditioned on such representations.

We show that the multimodal representations learned through M3L result in better sample efficiency and stronger generalization capabilities compared to settings that treat each modality separately. In particular, M3L demonstrates better zero-shot generalization to unseen objects and variations of the task scene, exploiting the representation power provided by multimodal reasoning. Moreover, we observe that the aforementioned generalization capabilities are substantially retained even if the representation encoder, trained with multimodal data, is deployed to a vision-only policy. This suggests that the generalization benefits of touch are strongly intertwined with how the policy learns its representations, offering the possibility to trade off a limited loss of performance with the additional complications of using touch sensors for robot deployment in the real world.

II. Related Work

Reinforcement Learning for Manipulation. The growing application of computer vision in robotics has enabled robotic manipulation policies trained from raw pixel observations through reinforcement learning (RL) [4], [5]. However, a vision-based policy struggles with occlusions and only enables a delayed response for contact-rich tasks.

Thanks to the recent advances in the development of high-resolution tactile sensors [10], [11], [12], [13], [14], [15], touch-based manipulation has tried to address the visual occlusion problem and enable reactive contact-rich manipulation with local information from tactile sensors. Common examples are in-hand manipulation [16], [17], Braille alphabet reading [18], pendulum swingups using learned feedforward [19] or feedback policies [20], or peg-hole insertion using primitive trajectories [21], [22] and task-specific controllers [23]. However, these approaches lack the use of visual information [24], [25], which is often required for global reasoning about the task.

The combination of vision and contact feedback has been investigated in various settings, such as model predictive control [26] and behavioral cloning [27]. More recently,

[1] UC Berkeley, please send correspondance to Carmelo Sferrazza. Email address: `csferrazza@berkeley.edu`

[2] KAIST

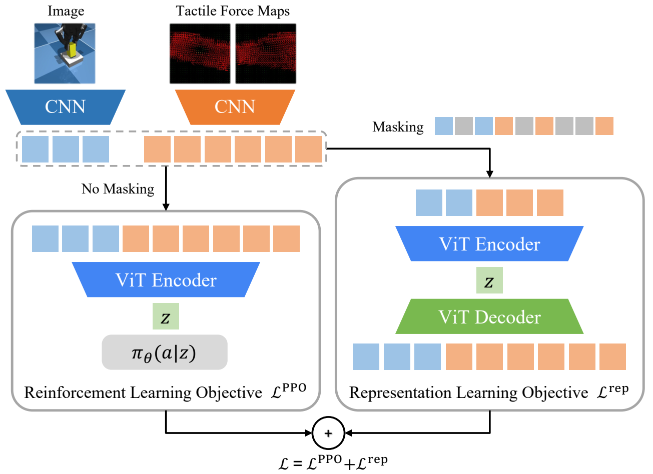


Image
Tactile Force Maps
CNN
CNN
Masking
No Masking
ViT Encoder
z
$\pi_\theta(a|z)$
Reinforcement Learning Objective $\mathcal{L}^{\text{PPO}}$
ViT Encoder
z
ViT Decoder
Representation Learning Objective $\mathcal{L}^{\text{rep}}$
+
$\mathcal{L} = \mathcal{L}^{\text{PPO}} + \mathcal{L}^{\text{rep}}$

Fig. 1: Masked Multimodal Learning (M3L) framework. M3L simultaneously optimizes a representation learning loss and a reinforcement learning objective. A policy is trained using Proximal Policy Optimization (PPO) [9], conditioned on multimodal representations learned through a masked autoencoder (MAE) [8]. By attending to each other within a unified vision transformer (ViT) encoder, visual and tactile data provide representations that lead to more generalizable policy learning. Note that the ViT encoders used for representation and policy learning share weights with each other.

various efforts have also been made in the context of model-free RL [28], [29], [30], which is the focus of our work and promises to learn control policies without the need for models of the system at hand or expert demonstrations. An end-to-end RL strategy from visual and tactile data, pre-processed through two separate neural networks, was shown in [31] on the Robosuite benchmark, where tactile signals were obtained through an approximation of the object depth map. In our work, rather than learning end-to-end, we focus on sample efficiency and generalization through a self-supervised representation learning objective.

Representation Learning for Manipulation. Representation learning has played a key role in reducing sample complexity when applying RL to high-dimensional observation spaces [32], [33], [34], [35]. In this context, several studies have focused on extracting condensed representations from tactile inputs [36], [37], [38]. A notable exception is [39], where a force-torque sensor was used in combination with vision, and a self-supervised learning architecture was found to improve sample efficiency compared to learning from raw data.

More recently, representation learning has been applied from visual and tactile data in contexts different from RL. Specifically, [40] trained tactile and visual encoders in a self-supervised manner, and exploited the extracted representations via imitation and residual learning [41] for manipulation tasks. In [42], a general perception module was proposed by training two separate (vision and touch) encoders using a contrastive approach. Our work differs in that we focus on fusing visual and high-resolution tactile inputs through a joint encoder that learns interrelations between the two modalities, particularly enhancing generalization capabilities. We focus on a specific class of representation learning algorithms, based on masked autoencoding [8].

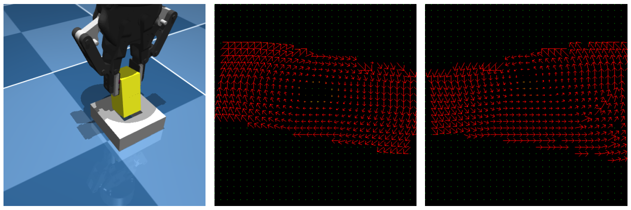
Masked Autoencoders for Manipulation. The idea of learning representations by reconstructing the masked parts of images [43], [44] has recently been scaled up inspired by the idea of masked language modeling in the language domain [45] and the introduction of the Transformer architecture [46]. Notably, [8] introduced Masked Autoencoders (MAE) that randomly mask patches of images and reconstruct the masked parts based on the vision transformer (ViT) architecture [47]. Recent works have demonstrated that MAE representations can be useful for learning manipulation policies from pixel observations [34], [35], [48], [33], [49]. In particular, the works closely related to ours have proposed to learn joint representations with MAEs and utilize it for robotic manipulation. For instance, [49] utilized frozen representations from a pre-trained vision-language multimodal MAE [50] for learning instruction-following manipulation policies. [48] trained an MAE with visual observations from multiple cameras and utilized it for RL. In this context, our work further demonstrates that learning joint vision-touch representations by training a multimodal MAE improves the sample efficiency and generalization of robotic manipulation policies.

III. Background

Reinforcement Learning (RL). We formulate the problem as a Markov decision process (MDP) [51], which is defined as a tuple $(\mathcal{S}, \mathcal{A}, p, r, \gamma)$. Here, $\mathcal{S}$ denotes the state space, $\mathcal{A}$ denotes the action space, $p(s_{t+1}|s_t, a_t)$ is the transition dynamics, r is the extrinsic reward function $r_t = r(s_t, a_t)$, and $\gamma \in [0, 1]$ is the discount factor. The goal of RL is to train a policy π to maximize the expected return. Our approach is compatible with any RL algorithm, but here we use Proximal Policy Optimization (PPO) [9] as our underlying RL algorithm due to its simplicity and scalability with parallel environments [52]. We refer to the online appendix[1] for more details about PPO.

Masked Autoencoding for Representation Learning. Masked autoencoding [8] is a self-supervised learning method that learns image representations by reconstructing the masked parts of images given the unmasked parts. Specifically, a masked autoencoder (MAE) first divides the images into non-overlapping square patches and adds positional embeddings [46] to the patches. Then, it randomly masks the patches, and a vision transformer (ViT) [47] encoder computes the visual embeddings of the remaining (unmasked) patches through a series of transformer layers [46]. Because the ViT encoder only processes a small subset of full patches (e.g., typically 25%), training becomes more compute-efficient and scalable. For decoding, learnable mask tokens [45] are concatenated with the unmasked patch representations and the positional embeddings are added in order to represent the position of masked patches to be reconstructed. Finally, a ViT decoder takes the concatenated inputs and outputs predicted pixel patches. All model parameters are

[1] `https://sferrazza.cc/m3l_site`



(a) Image (b) Tactile left (c) Tactile right

Fig. 2: Visualization of observations from the tactile insertion environment: (a) 64×64 visual input and (b, c) two 32×32 tactile inputs (taxels), where the color of the arrows indicates pressure (red means high pressure) and the direction indicates shear, following the convention in [53].

updated to minimize the mean squared error (MSE) between the predicted pixel patches and the original patches.

High-resolution Tactile Sensing Measurements. Modern high-resolution tactile sensors [54], [55], [56] may provide touch feedback in the form of spatially distributed quantities, such as deformation fields, strain fields, and force maps. In particular, force maps have been shown to be measurable in the real world through a variety of tactile sensors [57], [58], [59], are readily available through physics simulators (e.g., MuJoCo touch grid or the approach presented in [53]), and have been demonstrated as a valid abstraction to achieve successful sim-to-real transfer in highly dynamic manipulation tasks [20]. The elements comprising a force map are generally denoted as "taxels", i.e., the tactile dual of pixels. Such maps are often represented in a similar way as images, that is, in a `channels` × `height` × `width` form, where the channels are usually the three components: two for shear and one for pressure of the contact force, as shown in Figure 2.

IV. Method

In this section, we present **M**asked **M**ulti**m**odal **L**earning (**M3L**), a representation learning technique for reinforcement learning that targets robotic manipulation systems provided with vision and high-resolution touch. Specifically, M3L learns a policy conditioned on multimodal representations, which are extracted from visual and tactile data through a shared representation encoder. As illustrated in Figure 1, the M3L representations are trained by optimizing at the same time representation learning and reinforcement learning objectives:

$$\mathcal{L} = \mathcal{L}^{\texttt{rep}} + \mathcal{L}^{\texttt{PPO}}, \tag{1}$$

where $\mathcal{L}^{\texttt{rep}}$ is the multimodal representation learning objective (Section IV-A) and $\mathcal{L}^{\texttt{PPO}}$ is PPO's reinforcement learning objective (Section IV-B).

A. Representation Learning

M3L achieves multimodal representation learning by using both image and tactile data to update an MAE that learns to reconstruct both pixels and taxels at the same time. This can be written as following:

$$\mathcal{L}^{\texttt{rep}} = \text{MSE}^{\texttt{pixels}} + \beta_T \cdot \text{MSE}^{\texttt{taxels}}, \tag{2}$$

where β_T is a hyperparameter that balances the two MSE losses for vision and touch.

Note that as opposed to other representation learning approaches, such as contrastive learning, MAEs do not need discovering new data augmentations and invariances to design positives and negatives. On the other hand, patching and reconstructing in MAEs seamlessly apply to tactile data. In addition, the transformer architecture and the masking scheme support input sequences of variable length and facilitate design strategies particularly suited for multimodal data, e.g., vision and touch.

We list below the most relevant implementation details of our representation learning framework:

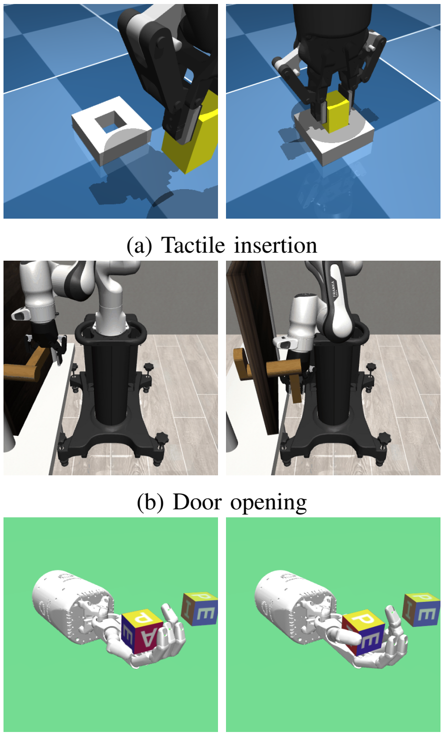
Early Convolutions. The MAE encoder has two preprocessing convolutional neural networks (CNNs) that compute convolutional features from pixels and taxels, respectively. Such convolutional features are then masked in place of the raw input patches. These early convolution layers help capturing small details in reconstruction [33].

Positional and Modality Embeddings. As standard for transformers, we add 2D sin-cos positional embeddings [60] to both the encoder and decoder features. In addition, we also add learnable 1D modality embeddings representing either visual or tactile streams, following the implementation of [50] for vision and language.

Reconstruction Pipeline. Convolutional features are computed as described above for k frames concatenated over time. In particular, we concatenate the frames in the channel dimension (e.g., concatenation of RGB images results in a $3k$-channel tensor). Frame stacking turned out to be crucial for success on the environments considered in our experiments (see Section VI). We then uniformly mask across visual and tactile features. Finally, we feed the unmasked convolutional features from both vision and touch into the MAE for reconstruction, so that the ViT encoder can attend to both modalities.

B. Policy Learning with M3L

The policy learning closely follows PPO with the exception of how the observations are extracted from the raw input data. At each time step, the image and tactile data are fed into the preprocessing CNNs. The CNN features are then added to the positional and modality embeddings and processed through the MAE encoder, without applying any masking. The extracted multimodal embeddings are then provided to the actor and critic networks. Each of these consist of a transformer layer that processes the embeddings and aggregates them through a global average pooling layer, and a multilayer perceptron (MLP) that outputs either the value (for the critic) or the mean of the action distribution (for the actor). Note that the gradients computed through the PPO loss are also propagated up to the MAE encoder and the CNNs. As a result, the CNNs and MAE encoder are updated to simultaneously optimize both representation and task learning. The overview of M3L is illustrated in Figure 1.



(a) Tactile insertion

(b) Door opening

(c) In-hand cube rotation

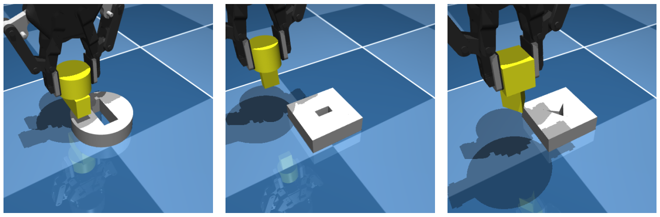
Fig. 3: We evaluate M3L on three simulated environments: (a) Tactile insertion, (b) Door opening, and (c) In-hand cube rotation. For each task, the left images represent initial configurations and the right images show task completion.

V. Simulation Environments

We perform our experiments in three simulated environments using MuJoCo [61]'s touch-grid sensor plugin, which aggregates contact forces into taxels. To the best of our knowledge, the following are the first examples where high-resolution force fields have been included in a MuJoCo robotics environment, which can also seamlessly render visual information. We will make our environments public for reproducibility and further research in visual-tactile manipulation.

A. Tactile Insertion

The tactile insertion environment consists of a peg object and a target frame where the peg can be tightly inserted, and the Menagerie's [62] Robotiq 2F-85 parallel-jaw gripper model, as shown in Figure 3a. Each finger has the silicone pad modeled as a rectangular prism (`box` geometry in MuJoCo). In MuJoCo, contact sensing with a box primitive is computed only at the four vertices. Therefore, we split the collision mesh of the box into a grid of smaller boxes, to increase the number of candidate contact points, and consequently the effective resolution of the force map. The resulting tactile observation is in the form of two 32×32 taxel maps (one per finger). Each taxel corresponds to a 3D force vector, which represents both shear and pressure forces, as shown in Figure 2. In addition, the observation also includes a 64×64 image.



(a) Training peg (b) Rectangle peg (c) V-shape peg

Fig. 4: We use 18 different training peg shapes, and 2 novel pegs (rectangle and V-shape) to test generalization on the tactile insertion task.

Each episode starts with the peg held between the gripper fingers with a randomized initial position of the gripper in the 3D space. We also randomize the shape of the peg (see all 18 peg shapes in the online appendix[1]), the shape of the target frame (square or circle), and the target hole location. The control inputs to the system are the 3D coordinates of the floating gripper, while we fix both the gripping force and gripper rotation. The task comprises 300 steps and is considered to be solved once the object position is within a small threshold from the target position. We use a dense reward, which is the negative distance between the peg and the target position, as well as a sparse task completion reward of 1000.

Note that while [22], [53] also address the insertion task with tactile information, they heavily rely on prior knowledge (e.g. initial estimate of the insertion region and open-loop insertion trajectory) and only learn to correct errors online using tactile information. On the other hand, our goal is to benchmark a general RL approach that can utilize vision and touch together to generate raw control actions, without requiring any prior information.

B. Door Opening

The door opening task from Robosuite [63] requires to open a locked door by turning the door handle and then pulling the door with a Franka robot arm and a Robotiq 2F-85 gripper, as shown in Figure 3b. We extend this environment by adding tactile sensors to gripper fingers as in the tactile insertion environment. The observation space comprises a 64×64 camera image and two 32×32 tactile maps. The action space consists of 3D delta end-effector position and rotation. Note that the gripper is always closed, holding the door handle.

To make the exploration problem easier and focus on generalizable skill learning, we provide additional dense rewards for opening the door and a sparse success reward of 300 when the door is opened. We initialize the robot to hold the door handle and the position of the door is fixed at $(0.07, 0.00)$. Each episode lasts for 300 steps but terminates when the door is opened or the gripper detaches from the door handle. To test generalization capability, we randomly initialize the door position, $x \sim [0.06, 0.10], y \sim [-0.01, 0.01]$ and use 10× higher friction and damping coefficients for hinges of both the door and door handle during testing.

C. In-hand Rotation

The in-hand cube rotation task is based on the in-hand block reorientation environment [64] provided through Gymnasium-Robotics. The environment relies on a Shadow Robot Dexterous Hand with 20D actions. We augment the visual observation with high-resolution force maps. Specifically, we add 3×3 force maps to each of the finger phalanges and to the palm of the hand. Through the use of zero-padding, we rearrange such force readings into a 32×32 map, as illustrated in the the online appendix[1].

The task consists in reorienting a colored cube to a predefined configuration, overlaid next to the actual hand-cube system (see Figure 3c). We use a reward of 100 when the cube is within a threshold from the target, in addition to the dense reward implemented in the original environment. To test generalization, we double the mass of the cube and slightly perturb the camera pose, and attempt the same reorientation task.

We found that this task requires a higher level of accuracy in the representations (e.g., to properly detect the different faces of the cube) compared to the previous two. For this reason, rather than directly optimizing the sum of two objectives in Equation (1), we perform a reinforcement learning gradient descent step every n representation learning gradient steps. In particular, given an RL batch size B, we split this batch in n chunks for the representation learning phase and then use the full batch for the reinforcement learning phase.

We present additional in-hand rotation tasks with different objects, e.g., an egg and a pen, in the online appendix[1].

VI. Experiments

In this section, we study the advantages of M3L in visual-tactile manipulation compared to baselines, and explain our design choices. In particular, we aim to answer the following questions:

- Does our multimodal approach improve generalization when manipulating unseen objects or dealing with scene variations?
- Is the representation learning loss beneficial compared to training the same architecture end-to-end via PPO?
- Can representations learned in a multimodal setting benefit vision-only deployment?
- Does attention across vision and touch lead to better overall performance?

A. Compared Methods

We compare the following approaches:

- **M3L:** our approach jointly learns visual-tactile representations using a multimodal MAE and the policy using PPO.
- **M3L (vision policy):** while representations are trained from both visual and tactile data, the policy takes only visual data, exploiting the variable input length of the ViT encoder.
- **Sequential:** an M3L architecture trained independently for the different modalities in sequence. At each MAE training iteration, we first propagate the gradient for vision and then for touch. In this way, visual features cannot attend tactile features and vice versa.
- **Vision-only (w/ MAE):** an MAE approach with the same architecture as M3L, but trained only from visual inputs.
- **End-to-end:** a baseline that trains the policy end-to-end but with the same encoder architecture as M3L.

B. Generalization Experiments

To evaluate the capabilities unlocked by multimodality, in this work we considered scenarios where both modalities are informative during most of the training episodes, i.e., visual information is most of the times sufficient to learn the task. Such a setting is especially suitable to isolate the effect of the multimodal representations (compared, for example, to the use of a single modality). In particular, we investigate the generalization capabilities unlocked by the multimodal representations when dealing with unseen objects or conditions. For the tactile insertion, we pretrain a policy on the set of 18 training objects, and test the zero-shot generalization on two different objects, which are a rectangular prism and V-shaped object (see Figure 4). Such objects are not seen during training, and the V-shaped object considerably differs from the training objects. For the door opening task, we randomize the initial position of the door, as well as the friction and damping coefficients of the hinges as described in Section V-B. All of these parameters were instead fixed during training. Finally, for the in-hand rotation, we double the mass of the cube and slightly perturb the camera pose.

The results are shown in Figure 5, with M3L consistently competing with or outperforming the end-to-end baseline and all the other representation learning approaches on all tasks. In particular, M3L substantially outperforms the vision-only approach, exploiting the power of multimodal representations. While the sequential baselines is competitive with M3L on the tactile insertion and in-hand reorientation tasks, it performs considerably worse on the door opening task. In particular, sequential training largely degrades due to observed training instabilities (see Figure 6), indicating that attention across modalities enables the extraction of stronger and more general representations.

Interestingly, we observe a considerable improvement of M3L with vision policy over the vision-only baseline. They key insight is that using touch only for training the representation encoder is sufficient to substantially fill the gap with M3L on all tasks. This opens several remarkable opportunities, namely, I) a limited loss of generalization performance when touch is used at training time, but removed at deployment time, II) the possibility of training multimodal representations exclusively in simulation, and transferring a stronger vision policy to the real-world, wherever visual sim-to-real transfer is achievable [65].

C. Training Performance

We report the learning curves for each task in Figure 6. The methods based on representation learning typically

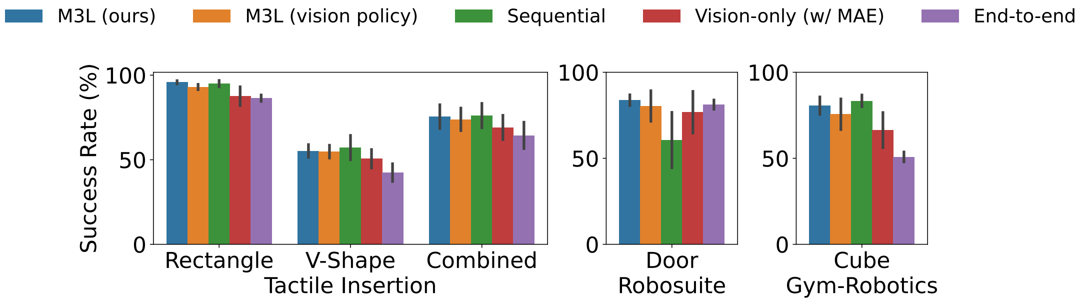
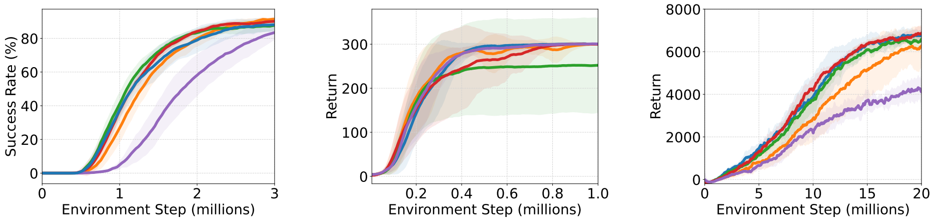


Fig. 5: Zero-shot generalization experiments on the three tasks. The bar plots show mean and standard error on 5 seeds, with 25 episodes run after training for the 4 last checkpoints on each seed.



(a) Tactile Insertion

(b) Door Opening

(c) Cube Rotation

Fig. 6: Learning curves investigating the advantages of M3L against baselines.

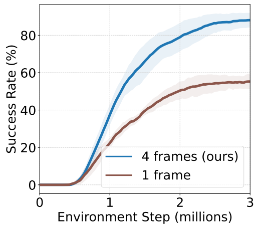


Fig. 7: Frame stacking ablation

exhibit higher sample efficiency compared to the end-to-end baseline, by exploiting the unsupervised reconstruction objective during training. M3L is the only approach that consistently achieves best in-task performance across the three tasks.

Finally, Figure 7 ablates the number of frames stacked together as an input to M3L (as explained in Section IV-A) for the tactile insertion task. The result may look counter-intuitive, given that the gripper is position controlled, and a single frame may appear sufficient to extract full information about the task. However, we hypothesize that when the framework is conditioned on a single frame, the encoder may struggle with visual occlusions. More importantly, contact information becomes much more relevant when stacking multiple frames, which act as a memory of a recent contact event. On the contrary, a single frame only signals current contact information, which immediately vanishes a step away from contact. Note that 4 frames are used as input to all the baselines considered in Figure 5 and Figure 6. Additional baselines and experiments are described in the online appendix[1], where we compare our representations against MVP [34], CLIP [66], and other related approaches [30].

VII. Conclusion

We have presented a systematic representation learning approach, Masked Multimodal Learning (M3L), to fuse visual and tactile data when using reinforcement learning for manipulation tasks. The results indicate that in addition to being sample efficient compared to an end-to-end baseline, the multimodal representations improve generalization to unseen objects and conditions over a variety of baselines. We notably observed how the benefits of training multimodal representations is partly retained when the representation encoder is applied to a vision policy. Finally, while contributing to tasks that cannot be solved with vision alone is certainly an important application of tactile sensing, this work indicates that touch can considerably contribute to efficient and generalizable manipulation also for tasks where vision appears to be sufficient. Therefore, we hope that this work opens new perspectives to incorporate this modality in a wider range of applications and learning frameworks.

Limitations and Future Work. Our method suffers from some of PPO's drawbacks, e.g., higher sample complexity compared to off-policy algorithms and struggle with difficult exploration problems. However, the modularity of the representation learning block makes it possible to combine it with other RL algorithms, and this will be the subject of future work.

An additional limitation of our approach is that it uses tactile data at all times, even when such data are uninformative, e.g., when contact is not taking place, which can potentially lead to slowing down learning. This information sparsity has been investigated in the past and a plausible solution indicated as tactile gating [31] may also be applied to our method.

Previous work that only relied on visual data [34] leveraged MAEs in a pretraining fashion, with a large encoder trained off-domain and directly deployed for learning a variety of tasks. Part of this success is due to the large availability of image and video datasets available to the research community [67], [68]. This is in contrast to the scarce availability of tactile datasets, often challenging to collect, especially when paired vision-touch data are required, such as for our approach. An interesting research direction would be to investigate how to leverage the large amount of available image data while only requiring a smaller portion of paired vision-touch data in a pretraining-finetuning fashion.

Considerations for Real-World Application. The current results were presented in simulation environments, which allowed us to thoroughly analyze and compare a wide range of architectural choices in a scalable manner. However, real-world applications may largely benefit from the findings of this work. Specifically, our algorithm shows improvements in sample efficiency compared to PPO from raw inputs (see Figure 6). Sample efficiency, together with the generalization properties showed by our approach, mark a crucial step towards the application of reinforcement learning on real-world robots, where we want to minimize both sample collection and retraining for each modification of the training task. Additionally, the performance benefits of using an M3L representation encoder for vision policies renders the possibility to train such policies in simulation with the availability of tactile signals, enabling the transfer of stronger vision policies to the real world, e.g., through the use of visual domain randomization.

Finally, the potential benefits of our work to real-world applications are confirmed by the successful transfer of approaches based on masked autoencoding from simulation to real-world systems in [48], [35]. In addition, the choice of force maps as tactile inputs has also proved its efficacy in sim-to-real transfer, as detailed in [20], [53].

ACKNOWLEDGMENT

We would like to thank Yuval Tassa and Kevin Zakka for the support with the touch grid sensor in MuJoCo. This work was supported in part by the SNSF Postdoc Mobility Fellowship 211086, the ONR MURI N00014-22-1-2773, the Office of Naval Research grant N00014-21-1-2769, Komatsu, InnoHK Centre for Logistics Robotics, BAIR Industrial Consortium, and an ONR DURIP grant.

REFERENCES

[1] R. Blake, K. V. Sobel, and T. W. James, "Neural synergy between kinetic vision and touch," *Psychological science*, vol. 15, no. 6, pp. 397–402, 2004.

[2] W. Zhou, Y. Jiang, S. He, and D. Chen, "Olfaction modulates visual perception in binocular rivalry," *Current Biology*, vol. 20, no. 15, pp. 1356–1358, 2010.

[3] E. Macaluso and J. Driver, "Spatial attention and crossmodal interactions between vision and touch," *Neuropsychologia*, vol. 39, no. 12, pp. 1304–1316, 2001.

[4] F. Zhang, J. Leitner, M. Milford, B. Upcroft, and P. Corke, "Towards vision-based deep reinforcement learning for robotic motion control," *arXiv preprint arXiv:1511.03791*, 2015.

[5] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke *et al.*, "Scalable deep reinforcement learning for vision-based robotic manipulation," in *Conference on Robot Learning*. PMLR, 2018, pp. 651–673.

[6] F. R. Hogan, J. Ballester, S. Dong, and A. Rodriguez, "Tactile dexterity: Manipulation primitives with tactile feedback," in *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2020, pp. 8863–8869.

[7] Y. She, S. Wang, S. Dong, N. Sunil, A. Rodriguez, and E. Adelson, "Cable manipulation with a tactile-reactive gripper," *The International Journal of Robotics Research*, vol. 40, no. 12-14, pp. 1385–1401, 2021.

[8] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.

[9] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[10] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, p. 2762, 2017.

[11] B. Ward-Cherrier, N. Pestell, L. Cramphorn, B. Winstone, M. E. Giannaccini, J. Rossiter, and N. F. Lepora, "The tactip family: Soft optical tactile sensors with 3d-printed biomimetic morphologies," *Soft robotics*, vol. 5, no. 2, pp. 216–227, 2018.

[12] C. Sferrazza and R. D'Andrea, "Design, motivation and evaluation of a full-resolution optical tactile sensor," *Sensors*, vol. 19, no. 4, p. 928, 2019.

[13] K. Park, H. Yuk, M. Yang, J. Cho, H. Lee, and J. Kim, "A biomimetic elastomeric robot skin using electrical impedance and acoustic tomography for tactile sensing," *Science Robotics*, vol. 7, no. 67, p. eabm7187, 2022.

[14] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer *et al.*, "Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 3838–3845, 2020.

[15] R. Bhirangi, T. Hellebrekers, C. Majidi, and A. Gupta, "Reskin:versatile, replaceable, lasting tactile skins," in *CoRL*, 2021.

[16] A. Melnik, L. Lach, M. Plappert, T. Korthals, R. Haschke, and H. Ritter, "Using tactile sensing to improve the sample efficiency and performance of deep deterministic policy gradients for simulated in-hand manipulation tasks," *Frontiers in Robotics and AI*, vol. 8, p. 538773, 2021.

[17] Z.-H. Yin, B. Huang, Y. Qin, Q. Chen, and X. Wang, "Rotating without seeing: Towards in-hand dexterity through touch," *Robotics: Science and Systems*, 2023.

[18] A. Church, J. Lloyd, R. Hadsell, and N. F. Lepora, "Deep reinforcement learning for tactile robotics: Learning to type on a braille keyboard," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 6145–6152, 2020.

[19] C. Wang, S. Wang, B. Romero, F. Veiga, and E. Adelson, "Swingbot: Learning physical features from in-hand tactile exploration for dynamic swing-up manipulation," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 5633–5640.

[20] T. Bi, C. Sferrazza, and R. D'Andrea, "Zero-shot sim-to-real transfer of tactile control policies for aggressive swing-up manipulation," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 5761–5768, 2021.

[21] S. Dong and A. Rodriguez, "Tactile-based insertion for dense box-packing," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 7953–7960.

[22] S. Dong, D. K. Jha, D. Romeres, S. Kim, D. Nikovski, and A. Rodriguez, "Tactile-rl for insertion: Generalization to objects of unknown geometry," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 6437–6443.

[23] S. Kim and A. Rodriguez, "Active extrinsic contact sensing: Application to general peg-in-hole insertion," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 10 241–10 247.

[24] Y. Lin, J. Lloyd, A. Church, and N. F. Lepora, "Tactile gym 2.0: Sim-to-real deep reinforcement learning for comparing low-cost high-resolution robot touch," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 754–10 761, 2022.

[25] S. H. Huang, M. Zambelli, J. Kay, M. F. Martins, Y. Tassa, P. M. Pilarski, and R. Hadsell, "Learning gentle object manipulation with curiosity-driven deep reinforcement learning," *arXiv preprint arXiv:1903.08542*, 2019.

[26] N. Fazeli, M. Oller, J. Wu, Z. Wu, J. B. Tenenbaum, and A. Rodriguez, "See, feel, act: Hierarchical learning for complex manipulation skills with multisensory fusion," *Science Robotics*, vol. 4, no. 26, p. eaav3123, 2019.

[27] H. Li, Y. Zhang, J. Zhu, S. Wang, M. A. Lee, H. Xu, E. Adelson, L. Fei-Fei, R. Gao, and J. Wu, "See, hear, and feel: Smart sensory fusion for robotic manipulation," in *Conference on Robot Learning*, 2022.

[28] L. Pecyna, S. Dong, and S. Luo, "Visual-tactile multimodality for following deformable linear objects using reinforcement learning," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 3987–3994.

[29] B. Wu, I. Akinola, J. Varley, and P. Allen, "Mat: Multi-fingered adaptive tactile grasping via deep reinforcement learning," in *Conference on Robot Learning*, 2019.

[30] H. Qi, B. Yi, Y. Ma, S. Suresh, M. Lambeta, R. Calandra, and J. Malik, "General In-Hand Object Rotation with Vision and Touch," in *Conference on Robot Learning (CoRL)*, 2023.

[31] J. Hansen, F. Hogan, D. Rivkin, D. Meger, M. Jenkin, and G. Dudek, "Visuotactile-rl: Learning multimodal manipulation policies with deep reinforcement learning," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 8298–8304.

[32] A. Zhan, R. Zhao, L. Pinto, P. Abbeel, and M. Laskin, "Learning visual robotic control efficiently with contrastive pre-training and data augmentation," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 4040–4047.

[33] Y. Seo, D. Hafner, H. Liu, F. Liu, S. James, K. Lee, and P. Abbeel, "Masked world models for visual control," in *Conference on Robot Learning*. PMLR, 2023, pp. 1332–1344.

[34] T. Xiao, I. Radosavovic, T. Darrell, and J. Malik, "Masked visual pre-training for motor control," *arXiv preprint arXiv:2203.06173*, 2022.

[35] I. Radosavovic, T. Xiao, S. James, P. Abbeel, J. Malik, and T. Darrell, "Real-world robot learning with masked visual pre-training," in *Conference on Robot Learning*. PMLR, 2023, pp. 416–426.

[36] Y. Chebotar, O. Kroemer, and J. Peters, "Learning robot tactile sensing for object manipulation," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 3368–3375.

[37] H. Van Hoof, N. Chen, M. Karl, P. van der Smagt, and J. Peters, "Stable reinforcement learning with autoencoders for tactile and visual data," in *2016 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2016, pp. 3928–3934.

[38] G. Sutanto, N. Ratliff, B. Sundaralingam, Y. Chebotar, Z. Su, A. Handa, and D. Fox, "Learning latent space dynamics for tactile servoing," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 3622–3628.

[39] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8943–8950.

[40] I. Guzey, B. Evans, S. Chintala, and L. Pinto, "Dexterity from touch: Self-supervised pre-training of tactile representations with robotic play," *arXiv preprint arXiv:2303.12076*, 2023.

[41] I. Guzey, Y. Dai, B. Evans, S. Chintala, and L. Pinto, "See to touch: Learning tactile dexterity through visual incentives," *arXiv preprint arXiv:2309.12300*, 2023.

[42] J. Kerr, H. Huang, A. Wilcox, R. Hoque, J. Ichnowski, R. Calandra, and K. Goldberg, "Learning self-supervised representations from vision and touch for active sliding perception of deformable surfaces," *arXiv preprint arXiv:2209.13042*, 2022.

[43] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, and L. Bottou, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion." *Journal of machine learning research*, vol. 11, no. 12, 2010.

[44] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2536–2544.

[45] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

[46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, vol. 30, 2017.

[47] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[48] Y. Seo, J. Kim, S. James, K. Lee, J. Shin, and P. Abbeel, "Multi-view masked world models for visual robotic manipulation," in *International Conference on Machine Learning*, 2023.

[49] H. Liu, L. Lee, K. Lee, and P. Abbeel, "Instruction-following agents with jointly pre-trained vision-language models," *arXiv preprint arXiv:2210.13431*, 2022.

[50] X. Geng, H. Liu, L. Lee, D. Schuurams, S. Levine, and P. Abbeel, "Multimodal masked autoencoders learn transferable representations," *arXiv preprint arXiv:2205.14204*, 2022.

[51] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT Press, 2018.

[52] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa *et al.*, "Isaac gym: High performance gpu-based physics simulation for robot learning," *arXiv preprint arXiv:2108.10470*, 2021.

[53] J. Xu, S. Kim, T. Chen, A. R. Garcia, P. Agrawal, W. Matusik, and S. Sueda, "Efficient tactile simulation with differentiability for robotic manipulation," in *Conference on Robot Learning*. PMLR, 2023, pp. 1488–1498.

[54] A. Yamaguchi and C. G. Atkeson, "Recent progress in tactile sensing and sensors for robotic manipulation: can we turn tactile sensing into vision?" *Advanced Robotics*, vol. 33, no. 14, pp. 661–673, 2019.

[55] A. C. Abad and A. Ranasinghe, "Visuotactile sensors with emphasis on gelsight sensor: A review," *IEEE Sensors Journal*, vol. 20, no. 14, pp. 7628–7638, 2020.

[56] K. Shimonomura, "Tactile image sensors employing camera: A review," *Sensors*, vol. 19, no. 18, p. 3933, 2019.

[57] C. Sferrazza and R. D'Andrea, "Sim-to-real for high-resolution optical tactile sensing: From images to three-dimensional contact force distributions," *Soft Robotics*, vol. 9, no. 5, pp. 926–937, 2022.

[58] L. Zhang, Y. Wang, and Y. Jiang, "Tac3d: A novel vision-based tactile sensor for measuring forces distribution and estimating friction coefficient distribution," *arXiv preprint arXiv:2202.06211*, 2022.

[59] D. Ma, E. Donlon, S. Dong, and A. Rodriguez, "Dense tactile force estimation using gelslim and inverse fem," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 5418–5424.

[60] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9640–9649.

[61] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 5026–5033.

[62] MuJoCo Menagerie Contributors, "MuJoCo Menagerie: A collection of high-quality simulation models for MuJoCo," 2022. [Online]. Available: http://github.com/deepmind/mujoco_menagerie

[63] Y. Zhu, J. Wong, A. Mandlekar, and R. Martín-Martín, "robosuite: A modular simulation framework and benchmark for robot learning," *arXiv preprint arXiv:2009.12293*, 2020.

[64] M. Plappert, M. Andrychowicz, A. Ray, B. McGrew, B. Baker, G. Powell, J. Schneider, J. Tobin, M. Chociej, P. Welinder, V. Kumar, and W. Zaremba, "Multi-goal reinforcement learning: Challenging robotics environments and request for research," *arXiv preprint arXiv:1802.09464*, 2018.

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

[66] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[67] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[68] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu *et al.*, "Ego4d: Around the world in 3,000 hours of egocentric video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 995–19 012.