M2CURL: Sample-Efficient Multimodal Reinforcement Learning via Self-Supervised Representation Learning for Robotic Manipulation

Fotios Lygerakis[1] Vedant Dave[1] Elmar Rueckert[1]

Abstract— One of the most critical aspects of multimodal Reinforcement Learning (RL) is the effective integration of different observation modalities. Having robust and accurate representations derived from these modalities is key to enhancing the robustness and sample efficiency of RL algorithms. However, learning representations in RL settings for visuotactile data poses significant challenges, particularly due to the high dimensionality of the data and the complexity involved in correlating visual and tactile inputs with the dynamic environment and task objectives. To address these challenges, we propose Multimodal Contrastive Unsupervised Reinforcement Learning (M2CURL). Our approach employs a novel multimodal self-supervised learning technique that learns efficient representations and contributes to faster convergence of RL algorithms. Our method is agnostic to the RL algorithm, thus enabling its integration with any available RL algorithm. We evaluate M2CURL on the Tactile Gym 2 simulator and we show that it significantly enhances the learning efficiency in different manipulation tasks. This is evidenced by faster convergence rates and higher cumulative rewards per episode, compared to standard RL algorithms without our representation learning approach. Project website: **https://sites.google.com/view/M2CURL/home**
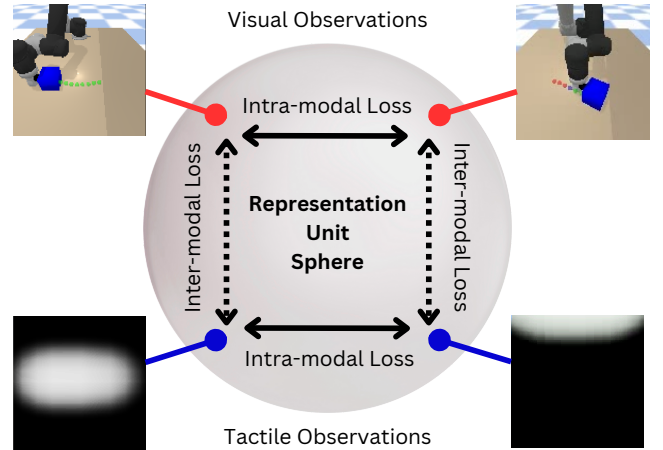
Fig. 1. Representation of M2CURL features on a unit sphere of codes. This diagram illustrates the projection of visual and tactile observations from high-dimensional space onto a unit sphere, where both intra and inter-modality losses are computed. These losses form the core of the contrastive multimodal loss, essential in M2CURL's learning process.

I. INTRODUCTION

Studies in cognitive science and psychology demonstrate that human beings possess the capacity to integrate diverse informational inputs, such as visual, auditory, and tactile data, for enhanced comprehension of their physical environment and informed decision-making processes [1–3]. The multidimensional nature of human information processing, emphasizing complexity and adaptability, mirrors the objectives of Multimodal Reinforcement Learning (MRL) [4–7]. MRL has gained prominence, with successful applications [8–11] underscoring its growing importance in the field.

Robust and accurate representations derived from these modalities are key to enhancing the robustness and sample efficiency of Reinforcement Learning (RL) algorithms. However, representing visuotactile data in RL settings is challenging due to the high dimensionality and complexity involved in correlating visual and tactile inputs with dynamic environments and task objectives [12, 13]. In tasks involving manipulation, vision is the primary sensory input for robots until they physically interact with an object or surface. Following this, tactile feedback becomes the main source of detailed information, especially in cases where the robot's arm obstructs the visual field. Thus, sophisticated learning techniques are essential to efficiently extract and integrate key features from visual and tactile data streams, ensuring consistent performance in unpredictable and rapidly changing environments. Furthermore, the intrinsic noise and variability present in sensory data, especially tactile sensing, add complexity to the process of learning representations [14, 15]. For RL algorithms, this is crucial, as the quality of representations directly influences the agent's decision-making capabilities and task success.

The integration of visual and tactile sensory information to address previously identified difficulties has become a primary focus of recent researches [16–20]. Advances in the realm of self-supervised learning (SSL) have been pivotal in this context, especially considering the limited availability of labeled datasets in tactile and visuotactile domains [21–24]. This approach has been proved to be highly beneficial for generating meaningful representations for RL applied to manipulation tasks, by reducing the dependency on labor-intensive labeled datasets, thus presenting a promising avenue to mitigate the hurdles in tactile data collection. Leveraging unlabeled data facilitates more efficient and robust representation learning, thereby enhancing the performance of RL algorithms in complex, real-world tasks. Despite the benefits, self-supervised learning approaches have not yet been effectively integrated with RL for executing complex manipulation tasks.

To tackle these challenges, we propose Multimodal Contrastive Unsupervised Reinforcement Learning (M2CURL).

[1]Chair of Cyber-Physical Systems, University of Leoben, Austria
*Contact Email: fotios.lygerakis@unileoben.ac.at

Our method facilitates the learning of intra and inter-modal representations by employing two pairs of encoders to compute four InfoNCE [25] losses. The within-modality (intra) losses, one for each modality, maximize the agreement among similar modality instances, while the cross-modality (inter) losses maximize the similarity between different modalities within the same sample. These representations are subsequently integrated into the RL algorithm, leading to accelerated convergence of the algorithm. Importantly, M2CURL is algorithm-agnostic, allowing for seamless integration into any pre-existing RL framework. We conducted experiments using the Tactile Gym 2 [26]. Our results confirm the M2CURL's effectiveness, revealing substantial improvements in learning efficiency and RL agent performance. Our findings indicate faster learning convergence and higher reward accumulation, compared to the baseline RL methods.

II. RELATED WORK

The fusion of visual and tactile modalities in robotics is an expanding research domain, with numerous studies exploring its complexities. This section aims to emphasize significant contributions made in this evolving field.

A. Unified Representation Learning for Visual and Tactile Modalities

Recent research underscores the importance of effectively integrating visual and tactile information to create efficient representations [27]. Li et al. [28] utilized cross-modal prediction, merging visual and tactile signals via conditional adversarial networks. They improved vision-touch interaction, prevented GAN mode collapse with data rebalancing, and included touch scale and location data in their model. Additionally, Lin et al. [29] learned to identify objects by a cross-modality instance recognition model. Similarly, Huaping et al. [16] devised a visual-tactile fusion framework utilizing a joint group kernel sparse coding approach to resolve the challenge of weak pairing in visual-tactile data samples. Luo et al.[18] learned a joint latent space shared by two modalities, i.e., vision and tactile data, for the task of cloth texture recognition. In the generative domain, Zhong et al. [30] used Neural Radiance Fields and Conditional GANs to generate camera-based tactile observations from desired poses. Yang et al. [31] used the latent diffusion model to learn representations that generate images from touch and vice-versa. Recently, Dave et al. [32] employed Multimodal Contrastive training to derive representations from both visual and tactile data, facilitating the performance of classification tasks across diverse datasets.

B. Multimodal sensory integration for Robotic Manipulation

Several advancements have been made in integrating multi-sensory information for improving grasping and manipulation tasks[33–36]. Calandra et al. [37, 38] found that incorporating tactile sensing into visual information significantly improves grasping outcomes. Lee et al. [39] incorporated self-supervised learning to derive representations from visual and tactile inputs. These representations were subsequently fused with optical flow and classical controllers for executing downstream manipulation tasks. Tian et al. [40] presented a tactile Model Predictive Control, which relies on a learned forward predictive model to execute goal-based actions. Chen et al. [21] developed the Visuo-Tactile Transformer (VTT), which combined visual and tactile data through spatial attention, showing improved efficiency in manipulation tasks. Kerr et al. [22] developed a self-supervised learning approach using intra-modal contrastive loss to learn representations for tasks like garment feature tracking and manipulation. However, this approach is not integrated with RL for action execution. Recently, Guzey et al. [24] demonstrated that applying self-supervised methods for learning tactile representations from a dataset of arbitrary, contact-rich interactions yielded enhanced outcomes in manipulation tasks.

III. BACKGROUND

This section provides a brief explanation of the core principles behind our M2CURL framework, focusing on mathematical models and referencing significant algorithms in the relevant literature.

A. Contrastive Learning

Contrastive Learning is a powerful technique in self-supervised learning that focuses on learning representations by distinguishing between similar (positive) and dissimilar (negative) pairs of samples. One of the most prominent contrastive losses in use is the InfoNCE loss [25]. Mathematically, it can be expressed as optimizing the following loss:

$$L = -\log \frac{\exp(\mathrm{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{N} \exp(\mathrm{sim}(z_i, z_k)/\tau)}$$

where z_i, z_j are representations of positive pairs, $\mathrm{sim}(\cdot)$ denotes a similarity measure (e.g., cosine similarity), τ is a temperature scaling parameter, and N is the number of negative samples.

Recent advancements in contrastive learning include approaches like SimCLR [41] and MoCo [42], which have set new benchmarks in unsupervised representation learning in image and language processing domains. These methods emphasize the importance of a rich set of augmentations and a large number of negative samples to learn generalizable features.

B. Soft Actor-Critic Algorithm

The Soft Actor-Critic (SAC) algorithm [43] is an off-policy actor-critic framework that incorporates entropy regularization to effectively balance exploration and exploitation. Being off-policy, SAC benefits from the ability to learn from past experiences stored in a replay buffer, enhancing sample efficiency. However, this approach may present some challenges in maintaining stability compared to on-policy methods like PPO [44], particularly when adapting to varying
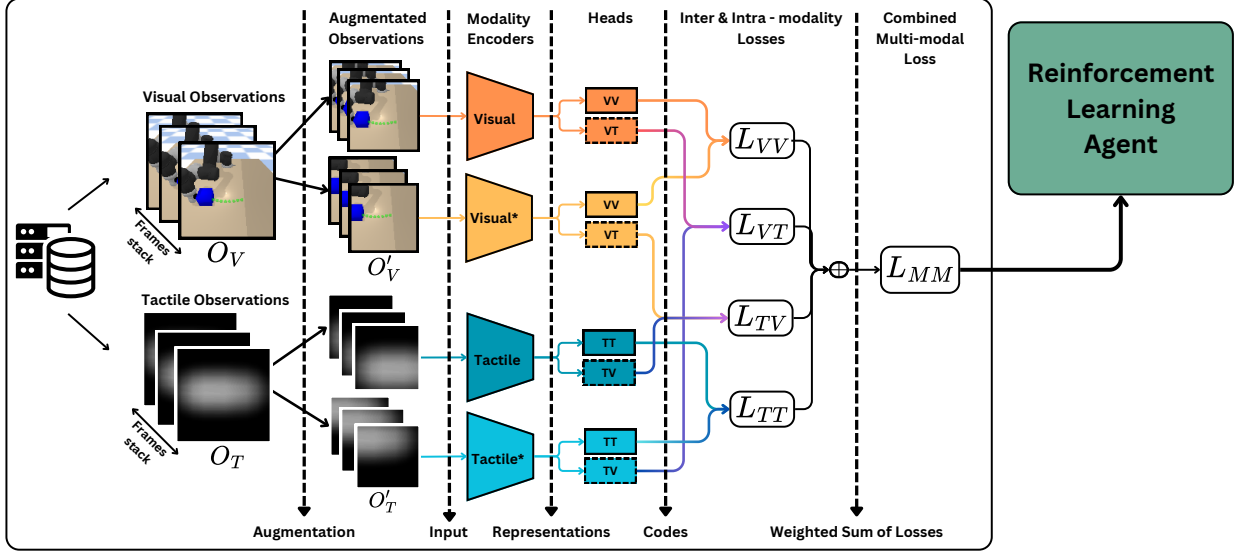
Fig. 2. The M2CURL Architecture: First, a batch of visuotactile observations are sampled from the replay buffer. Then, two random augmentations are applied for the query (online) and key (momentum) encoders, and their representation is computed. The query and key representations are used to compute the inter and intra-modality codes using the respective heads, from which the different inter and intra-modality losses are computed. Finally, the weighted sum of the sub-losses is passed to the RL algorithm as a combined multimodal contrastive loss $\mathcal{L}_{MM}$. Momentum encoders are denoted with *.

environments and during the hyperparameter tuning process. The SAC's objective function is defined as:

$$J(\pi) = \mathbb{E}_{(s_t,a_t) \sim \rho_\pi} \left[\sum_t \gamma^t \left(R(s_t, a_t) + \alpha \mathcal{H}(\pi(\cdot|s_t)) \right) \right]$$

where ρ_π is the state-action distribution under policy π, $R(s_t, a_t)$ is the reward function, γ is the discount factor, α is the temperature parameter determining the importance of the entropy term $\mathcal{H}$, and $\pi(\cdot|s_t)$ is the policy. SAC's strength lies in its robustness and ability to handle high-dimensional action spaces, but it can be computationally intensive due to the need for frequent policy updates.

C. Proximal Policy Optimization

Proximal Policy Optimization (PPO) [44] is an on-policy algorithm that optimizes a modified version of the expected return to maintain a balance between policy improvement and stability. Being on-policy, PPO updates policies using data collected by the current policy, ensuring relevance and consistency of the learning process. However, this approach can be less sample-efficient compared to off-policy methods. The PPO objective function with clipping is:

$$L^{CLIP}(\theta) = \mathbb{E}_t \left[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_t) \right]$$

where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$ is the probability ratio of the new policy to the old policy, $\hat{A}_t$ is the advantage estimate at time t, and ε is a hyperparameter that controls the clipping range. PPO's main advantage is its stability and ease of implementation, but its on-policy nature requires more interactions with the environment, potentially making it slower to converge in complex tasks.

IV. MULTIMODAL CONTRASTIVE UNSUPERVISED REINFORCEMENT LEARNING (M2CURL)

The following sections describe the M2CURL method architecture, its novel visuotactile contrastive learning strategy, and its simple integration with off-the-shelf RL algorithms.

A. M2CURL Method Architecture

The M2CURL architecture is tailored for processing multimodal data, with a specific focus on visual ($\mathcal{O}_V$) and tactile ($\mathcal{O}_T$) observations. It comprises two main components: the encoders and the inter/intra modality heads.

Online Encoders: There are two online encoders, $f_{visual}(\theta_v)$ and $f_{tactile}(\theta_t)$. The online encoders are used to extract the features used by the RL algorithm. f_{visual} is responsible for processing visual data ($\mathcal{O}_V$), converting it into a feature representation z_v. Similarly, $f_{tactile}$ processes tactile data ($\mathcal{O}_T$), resulting in a tactile feature representation z_t.

Momentum Encoders: Each encoder has a corresponding momentum encoder, $m_{visual}(\theta_{v,m})$ and $m_{tactile}(\theta_{t,m})$, which are updated using a momentum-based approach from the online encoder. Their representations are solely used to compute the contrastive losses. The momentum encoding ensures stability and temporal consistency in the learned representations [42]. The momentum update rules are given by:

$$\theta_{v,m} \leftarrow \alpha\theta_{v,m} + (1-\alpha)\theta_v$$
$$\theta_{t,m} \leftarrow \alpha\theta_{t,m} + (1-\alpha)\theta_t$$

where α is the momentum coefficient, controlling the rate of update from the primary encoders to the momentum encoders.

Heads: Each encoder has two heads, each being a 2-layer neural network. For the visual encoder, the heads are:

- **Vision-to-vision head** (H_{vv}), which focuses on learning representations within the visual domain:

$$c_{vv}(z_v) = H_{vv}z_v, \quad \text{where} \quad z_v = f_{visual}(o_V)$$

- **Vision-to-tactile head** (H_{vt}), dedicated to correlating visual features with tactile data:

$$c_{vt}(z_v) = H_{vt}z_v$$

For the tactile encoder, the heads are:

- **Tactile-to-tactile head** (H_{tt}), concentrating on learning within the tactile domain:

$$c_{tt}(z_t) = H_{tt}z_t, \quad \text{where} \quad z_t = f_{tactile}(o_T)$$

- **Tactile-to-vision head** (H_{tv}), for integrating tactile features into the visual representation space:

$$c_{tv}(z_t) = H_{tv}z_t$$

These components collectively enable the encoders to be trained via a contrastive learning approach, thereby developing rich and effective multimodal representations. This enhances the RL agent's performance in environments requiring sophisticated sensory integration. The overall architecture, including the momentum encoders and the two-layer neural network heads, is illustrated in Figure 2.

B. Visuotactile Augmentation Strategy

In M2CURL, data augmentation is crucial for multimodal contrastive learning. We specifically utilize random cropping and normalization, as these methods have shown to be highly effective [45]. Random cropping ($\mathscr{A}_{crop}$) introduces variations in the input data by extracting different regions of the images, thereby improving the model's robustness to changes in perspective and scale. Normalization ($\mathscr{N}$), on the other hand, standardizes the pixel values across the dataset, which has been found to enhance learning stability and performance. The augmentation process for both visual (o_V) and tactile (o_T) observations is formulated as:

$$o'_V = \mathscr{N}(\mathscr{A}_{crop}(o_V)), \quad o'_T = \mathscr{N}(\mathscr{A}_{crop}(o_T))$$

where o'_V and o'_T are the augmented visual and tactile observations, respectively. This specific combination of augmentation techniques ensures a diverse range of perspectives and scales in the training data, which is key for learning invariant and robust features from visual data.

C. Multimodal Contrastive Loss

The M2CURL framework employs a sophisticated contrastive loss function that integrates both intra-modal and inter-modal learning, leveraging the specialized encoder heads for each modality.

Intra-modal Contrastive Learning: For intra-modal learning, contrastive loss is computed separately for each modality:

- *Vision-to-Vision Contrastive Loss ($\mathscr{L}_{VV}$):* This loss is calculated using the vision-to-vision head H_{vv}. It contrasts different augmented versions of the same visual

observation to enhance feature learning within the visual domain.

- *Tactile-to-Tactile Contrastive Loss ($\mathscr{L}_{TT}$):* Similarly, this loss uses the tactile-to-tactile head H_{tt} to contrast different tactile observations, fostering feature extraction in the tactile domain.

The loss for each modality is computed using the InfoNCE formula:

$$\mathscr{L}_{VV/TT} = -\log \frac{\exp(\text{sim}(c_{vv/tt}(z_i), c_{vv/tt}(z_j))/\tau)}{\sum_{k=1}^{N} \exp(\text{sim}(c_{vv/tt}(z_i), c_{vv/tt}(z_k))/\tau)}$$

Inter-modal Contrastive Learning: Inter-modal learning involves aligning the feature spaces across the visual and tactile modalities:

- *Vision-to-Tactile Contrastive Loss ($\mathscr{L}_{VT}$):* This part of the loss function uses the vision-to-tactile head H_{vt} to learn cross-modal representations, focusing on understanding the correlation between visual features and tactile sensory data.

- *Tactile-to-Vision Contrastive Loss ($\mathscr{L}_{TV}$):* Similarly, this loss employs the tactile-to-vision head H_{tv} for learning cross-modal representations, aimed at comprehending how tactile features correlate with visual information.

The inter-modal contrastive loss is also based on the InfoNCE loss formula, adapted for cross-modal comparisons:

$$\mathscr{L}_{VT/TV} = -\log \frac{\exp(\text{sim}(c_{vt/tv}(z_i), c_{tv/vt}(z_j))/\tau)}{\sum_{k=1}^{N} \exp(\text{sim}(c_{vt/tv}(z_i), c_{tv/vt}(z_k))/\tau)}$$

Combined Contrastive Loss: The combined multimodal contrastive loss in M2CURL, $\mathscr{L}_{MM}$, integrates the intra-modal and inter-modal losses, each modulated by a specific balancing coefficient. This formulation allows for tailored learning from both within-modality and across-modality representations.

The combined loss is defined as:

$$\mathscr{L}_{MM} = \lambda_{VV}\mathscr{L}_{VV} + \lambda_{TT}\mathscr{L}_{TT} + \lambda_{VT}\mathscr{L}_{VT} + \lambda_{TV}\mathscr{L}_{TV}$$

where λ_{VV}, λ_{TT}, λ_{VT}, and λ_{TV} are the coefficients for vision-to-vision, tactile-to-tactile, vision-to-tactile, and tactile-to-vision learning, respectively. These coefficients enable precise control over the learning process, ensuring a balanced approach to multimodal representation integration and enhancing RL agent performance in diverse sensory environments.

D. Integration with Reinforcement Learning

The M2CURL framework's integration with RL is versatile, accommodating various RL architectures, including those beyond the traditional actor-critic paradigm. The key component, the multimodal contrastive loss ($\mathscr{L}_{MM}$), is adaptable to the specific RL setup in use.

For actor-critic architectures, $\mathscr{L}_{MM}$ is integrated into the actor's loss function to enhance policy learning:

$$\mathscr{L}'_{actor} = \mathscr{L}_{actor} + \beta \mathscr{L}_{MM}$$

In this setup, $\mathscr{L}_{actor}$ denotes the standard RL loss of the actor, and β acts as a weighting factor for the multimodal contrastive loss. Additionally, in such architectures, the critic's encoder is periodically updated by copying the weights from the actor's encoder. This ensures consistency between the policy evaluation by the critic and the policy updates by the actor.

In cases where an actor-critic algorithm is not employed, the integration of $\mathscr{L}_{MM}$ adapts similarly to the specific algorithmic structure. For instance, the primary loss function of the RL algorithm can be modified to include $\mathscr{L}_{MM}$, thereby infusing the learning process with multimodal data insights. This flexibility allows M2CURL to enhance a wide range of RL approaches, improving learning efficiency and policy performance across diverse environments and tasks.

V. EXPERIMENTAL FRAMEWORK

A. Evaluation Metrics

In evaluating our methods and baselines, we focus on data efficiency and overall performance at two key milestones: 100k and 500k environment steps. This approach allows us to assess both the speed of initial learning and the asymptotic performance. We use two primary metrics for evaluation:

1) Sample Efficiency: Measured by the number of steps required by the baselines to match the performance of M2CURL at fixed environment steps (100k or 500k).
2) Performance: Assessed by comparing the mean cumulative reward per episode.

B. Environments

The experiments are conducted in three Tactile Gym 2.0 environments, each chosen for its distinct challenges and complexity, providing a comprehensive testing ground for our algorithms.

1) **Object Push**: This environment involves the task of pushing a cube object along a trajectory generated by OpenSimplex Noise, emphasizing the need for precise manipulation and adaptability to unpredictable paths.
2) **Edge Follow** : The task here is to traverse a flat edge randomly oriented within a 360-degree range. The challenge is in keeping the edge centered on the sensor, requiring high precision and adaptability.
3) **Surface Follow V2**): A verticalized version of the *Surface Follow V2* environment designed for training 4-Degree-of-Freedom (4-DoF) robots. This variant adds complexity by altering the orientation and dynamics of navigation, testing the robot's ability to adapt to vertical surfaces and different gravitational dynamics.

These environments are ideal for evaluating the performance of M2CURL)in robotic manipulation tasks. They offer a thorough assessment of the algorithms in terms of precision, complexity, and coordination of different modalities to complete the task.

C. Baselines for Benchmarking Sample Efficiency

In our benchmarking process, M2CURL's performance was evaluated alongside SAC and PPO, including their versions augmented with RAD(Random Augmentation for Data-efficiency) [46], a technique originally developed for visual tasks and here adapted for visuotactile perception. This comparative analysis aims to showcase the advancements M2CURL introduces in terms of learning convergence within tactile-rich environments. While comparing against the original versions of both RL algorithms, we want to rule out the efficacy that data augmentations may have on learning efficient representations of the visuotactile observations. Therefore, we implement the same augmentation approach used in M2CURL to preprocess the observations before feeding them into the RAD versions of the two RL algorithms. This comparison aims to explore whether RAD's augmentation techniques bolster SAC and PPO in multimodal environments and how M2CURL fares in harnessing visuotactile information for learning.

D. Implementation Details

M2CURL was implemented using Stable Baselines 3 [47] and was tested with a simulated UR5 robot equipped with a DIGIT tactile sensor, using TCP velocity control. For the *Object Push* and *Surface Follow V2* environments, we used random trajectories, and for the *Edge Follow* variable vertical distances. A dense reward structure was applied for effective learning feedback. The modality weights ($\lambda_{VV,TT,VT,TV}$) were uniformly set to 1, with the contrastive loss weight (β) adjusted differently for SAC (0.1) and PPO (1) to address overfitting issues observed in SAC due to sample diversity. A higher temperature parameter (τ) was used for SAC(0.1) than in PPO (0.05) to prevent overfitting of λ_{MM}. This choice was made to address SAC's limited sample diversity. This adjustment softens the distribution over sample pairs, ensuring that even less similar samples get some positive probability. Consequently, it reduces the stark contrast in scores between positive and negative pairs, helping to mitigate overfitting. Other hyperparameters and network architectures followed the Tactile Gym 2 simulator [26] settings, with heads comprising two fully connected layers with a hidden dimension of 2048.

VI. RESULTS

Our experimental evaluation assesses the performance of the M2CURL framework against the SAC and PPO frameworks, in their standard and RAD-augmented versions. We also compare against the same algorithms with access to the actual state of the robot instead of the visuotactile observations. The different milestones at 100K and 500K timesteps were chosen to understand the initial adaptation and the more mature phase of learning of each algorithm, providing insights into both the early learning phase (100k steps) and the more mature phase of learning (500k steps).

In the *Object Push* task, M2CURL demonstrated a significant advantage over the baselines. At 500k steps, M2CURL achieved higher scores for both SAC and PPO frameworks.

500K Steps	M2CURL SAC	RAD SAC	SAC	SAC-state	M2CURL PPO	RAD PPO	PPO	PPO-state
Object Push	**-63.3±4.3**	-94.0±12.5	-64.8±8.2	-83.1±12.6	**-67.5±2.0**	-162.0±3.3	-187.2±3.1	-18.8±1.7
Edge Follow	**-24.1±1.3**	-27.3±1.4	-27.2±4.5	-35.5±3.8	**-15.6±2.1**	-19.0±2.3	-21.6±4.6	-30.5±5.0
Surface Follow V2	**-37.4±2.1**	-46.8±5.1	-178.6±24.1	-39.7±5.01	**-48.0±2.3**	-64.2±2.8	-61.4±2.6	-8.9±1.5
100K Step Scores								
Object Push	**-114.3±13.4**	-181.9±21.9	-190.6±14.1	-111.6±13.5	**-179.7±11.6**	-183.4±19.8	-237.6±14.5	-52.3±4.2
Edge Follow	**-31.0±0.9**	-38.9±1.9	-42.8±2.8	-35.5±3.7	**-14.7±1.2**	-17.7±2.4	-32.6±1.8	-44.8±4.1
Surface Follow V2	**-45.1±3.2**	-78.6±6.1	-63.8±5.5	-60.3±4.6	**-6.5±0.7**	-23.7±1.1	-18.9±2.6	-16.6±1.4

Notably, M2CURL showed its rapid adaptation capability in the early learning phase, outperforming others at 100k steps, indicative of its enhanced sample efficiency. The effectiveness of the multimodal contrastive loss becomes more apparent by comparing the performance of the M2CURL PPO algorithm with the RAD SAC and SAC algorithms. In general, off-policy algorithms, like SAC, are more sample-efficient than on-policy algorithms, like PPO. This is apparent by comparing the performance of the two algorithms in Table I. Despite that, M2CURL PPO manages to converge faster than the RAD SAC and SAC, highlighting the effectiveness of our method. M2CURL's sample efficiency is consistent in the 'Edge Following' task too, outpacing both the standard and RAD-augmented versions of either RL framework.

TABLE II

ABLATION STUDY IN THE 'OBJECT-PUSH' ENVIRONMENT, COMPARING THE PERFORMANCE OF SAC WHEN TRAINED EXCLUSIVELY WITH INTRA-MODALITY LOSSES, EXCLUSIVELY WITH INTER-MODALITY LOSSES, AND THE ORIGINAL M2CURL LOSS.

Steps	M2CURL	Intra-Modality	Inter-Modality
500K	-63.37±6.4	-140.45±19.3	-63.73±8.7
100K	-114.32±16.6	-154.76±13.8	-261.40±45.2

In the *Surface Follow V2* task, M2CURL notably mitigated the divergence that occurred in the later stages of training the PPO instances. This task's complexity, requiring navigation and adaptation to vertical surfaces and varied dynamics, typically poses challenges for on-policy algorithms like PPO. However, the multimodal contrastive learning approach in M2CURL played a crucial role in ensuring learning stability and efficiency, demonstrating its robustness and adaptability in this challenging environment.

The results across both tasks and at both the early (100k) and later (500k) stages of learning highlight M2CURL's sample efficiency and long-term performance capabilities. The outcomes observed at the 100K mark are especially indicative, as they demonstrate M2CURL's ability to quickly grasp and adapt to new tasks, a crucial aspect in dynamic environments where rapid learning is essential. The consistent high-level performance of M2CURL in later stages, coupled with its ability to outperform even algorithms with direct access to the state, highlights the framework's proficiency in learning efficient visuotactile representations.
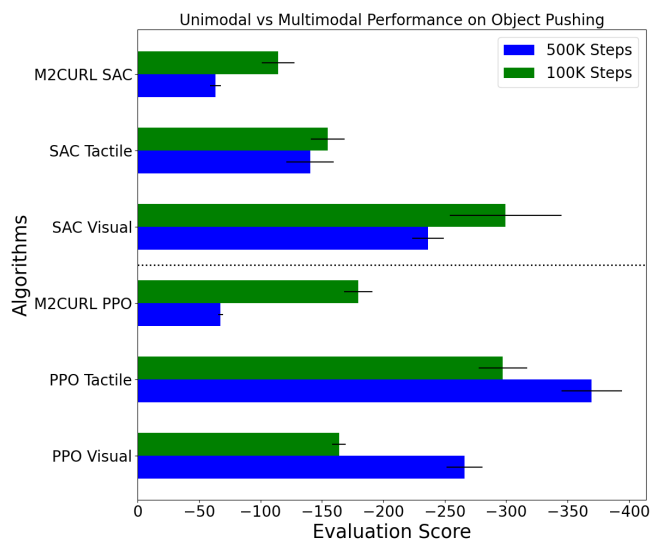

Fig. 3. Performance comparison of SAC and PPO algorithms using visual or tactile observations, against M2CURL using visuotactile observations.

To further assess M2CURL's performance relative to a simpler contrastive loss framework, we conducted ablation studies detailed in Table II. In these studies, the SAC algorithm was trained in the 'Object-Push' environment using only intra-modality losses ($\lambda_{VT} = \lambda_{TV} = 0$) and separately with only inter-modality losses ($\lambda_{VV} = \lambda_{TT} = 0$). The results, as presented in the table, unequivocally demonstrate that a combined approach of both intra and inter-modality losses within M2CURL leads to superior visuotactile representation learning, underscoring the framework's advanced capability.

Finally, we aimed to highlight the importance of modality fusion in tactile-rich settings. In our study shown in Figure 3, we compared unimodal (visual-only and tactile-only) reinforcement learning against the multimodal approach in M2CURL. The results clearly showed that M2CURL's integration of both visual and tactile data significantly outperforms unimodal methods for either RL algorithm. This underscores the effectiveness of multimodal learning in complex environments, demonstrating the advantage of combining modalities for improved reinforcement learning performance.

VII. Conclusions

In this paper, we introduced M2CURL, a novel framework that leverages a multimodal contrastive loss to enhance the efficiency of RL agents in tactile-rich robotic manipulation tasks. M2CURL processes both inter-modality and intra-modality losses, facilitating the learning of efficient representations. These representations are then utilized by the actor and critic components of two state-of-the-art RL algorithms, an on-policy and an off-policy one. Our empirical results underscore the substantial benefits M2CURL offers compared to the simple representation concatenation for the two modalities in RL algorithms. M2CURL's algorithm-agnostic nature further enhances its utility, allowing for easy integration with various existing RL algorithms. Our findings suggest that M2CURL can play a pivotal role in advancing robotic manipulation tasks, especially those requiring the integration of complex and varied sensory inputs. The framework's ability to rapidly adapt and maintain robust performance over time offers new opportunities for deploying more intelligent and capable robotic systems in a wide range of real-world applications. Future work will focus on extending the application of M2CURL to physical robotic systems and exploring its scalability and effectiveness in more diverse environments and with additional sensory modalities.

References

[1] Elizabeth Spelke. "Infants' intermodal perception of events". In: *Cognitive psychology* 8.4 (1976), pp. 553–560.

[2] Joseph W Sullivan and Frances D Horowitz. "Infant intermodal perception and maternal multimodal stimulation: Implications for language development." In: *Advances in infancy research* (1983).

[3] Arlene S Walker-Andrews. "Infants' perception of expressive behaviors: differentiation of multimodal information." In: *Psychological bulletin* 121.3 (1997), p. 437.

[4] Jiquan Ngiam et al. "Multimodal Deep Learning". In: Jan. 2011, pp. 689–696.

[5] Nitish Srivastava and Russ R Salakhutdinov. "Multimodal learning with deep boltzmann machines". In: *Advances in neural information processing systems* 25 (2012).

[6] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. "Multimodal Machine Learning: A Survey and Taxonomy". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.2 (2019), pp. 423–443. DOI: 10.1109/TPAMI.2018.2798607.

[7] Dana Lahat, Tülay Adali, and Christian Jutten. "Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects". In: *Proceedings of the IEEE* 103.9 (2015), pp. 1449–1477. DOI: 10.1109/JPROC.2015.2460697.

[8] Samira Ebrahimi Kahou et al. "Emonets: Multimodal deep learning approaches for emotion recognition in video". In: *Journal on Multimodal User Interfaces* 10 (2016), pp. 99–111.

[9] Jiaping Zhang, Tiancheng Zhao, and Zhou Yu. "Multimodal hierarchical reinforcement learning policy for task-oriented visual dialog". In: *arXiv preprint arXiv:1805.03257* (2018).

[10] Xin Qian, Ziyi Zhong, and Jieli Zhou. "Multimodal machine translation with reinforcement learning". In: *arXiv preprint arXiv:1805.02356* (2018).

[11] Johanna Hansen et al. "Visuotactile-RL: Learning Multimodal Manipulation Policies with Deep Reinforcement Learning". In: *2022 International Conference on Robotics and Automation (ICRA)*. 2022, pp. 8298–8304. DOI: 10.1109/ICRA46639.2022.9812019.

[12] Bin He et al. "Review of Bioinspired Vision-Tactile Fusion Perception (VTFP): From Humans to Humanoids". In: *IEEE Transactions on Medical Robotics and Bionics* 4.4 (2022), pp. 875–888. DOI: 10.1109/TMRB.2022.3215749.

[13] Shuo Gao, Yanning Dai, and Arokia Nathan. "Tactile and vision perception for intelligent humanoids". In: *Advanced Intelligent Systems* 4.2 (2022), p. 2100074.

[14] Zihao Ding et al. "Adaptive visual–tactile fusion recognition for robotic operation of multi-material system". In: *Frontiers in Neurorobotics* 17 (2023), p. 1181383.

[15] Yijiong Lin et al. "Attention for Robot Touch: Tactile Saliency Prediction for Robust Sim-to-Real Tactile Control". In: *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2023, pp. 10806–10812.

[16] Huaping Liu et al. "Visual–tactile fusion for object recognition". In: *IEEE Transactions on Automation Science and Engineering* 14.2 (2016), pp. 996–1008.

[17] Hao Li et al. "See, Hear, and Feel: Smart Sensory Fusion for Robotic Manipulation". In: *CoRL*. 2022.

[18] Shan Luo et al. "Vitac: Feature sharing between vision and tactile sensing for cloth texture recognition". In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2018, pp. 2722–2727.

[19] Zhiqin Zhu et al. "A novel multi-modality image fusion method based on image decomposition and sparse representation". In: *Information Sciences* 432 (2018), pp. 516–529.

[20] Sarmad Maqsood and Umer Javed. "Multi-modal medical image fusion based on two-scale image decomposition and sparse representation". In: *Biomedical Signal Processing and Control* 57 (2020), p. 101810.

[21] Yizhou Chen et al. "Visuo-Tactile Transformers for Manipulation". In: *6th Annual Conference on Robot Learning*. 2022. URL: https://openreview.net/forum?id=JqqSTgdQ85F.

[22] Justin Kerr et al. *Self-Supervised Visuo-Tactile Pre-training to Locate and Follow Garment Features.* 2023. arXiv: `2209.13042 [cs.RO]`.

[23] Fengyu Yang et al. "Touch and Go: Learning from Human-Collected Vision and Touch". In: *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track.* 2022.

[24] Irmak Guzey et al. "Dexterity from Touch: Self-Supervised Pre-Training of Tactile Representations with Robotic Play". In: *7th Annual Conference on Robot Learning.* 2023. URL: `https://openreview.net/forum?id=EXQ0eXtX3OW`.

[25] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. "Representation Learning with Contrastive Predictive Coding". In: *CoRR* abs/1807.03748 (2018). arXiv: `1807.03748`. URL: `http://arxiv.org/abs/1807.03748`.

[26] Yijiong Lin et al. "Tactile Gym 2.0: Sim-to-real Deep Reinforcement Learning for Comparing Low-cost High-Resolution Robot Touch". In: ed. by R. Liu A.Banerjee. Vol. 7. Proceedings of Machine Learning Research 4. IEEE, Aug. 2022, pp. 10754–10761. DOI: `10.1109/LRA.2022.3195195`.

[27] Nicolás Navarro-Guerrero et al. "Visuo-haptic object perception for robots: an overview". In: *Autonomous Robots* 47.4 (2023), pp. 377–403.

[28] Yunzhu Li et al. "Connecting touch and vision via cross-modal prediction". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2019, pp. 10609–10618.

[29] Justin Lin, Roberto Calandra, and Sergey Levine. "Learning to Identify Object Instances by Touch: Tactile Recognition via Multimodal Matching". In: *2019 International Conference on Robotics and Automation (ICRA).* 2019, pp. 3644–3650. DOI: `10.1109/ICRA.2019.8793885`.

[30] Shaohong Zhong et al. "Touching a NeRF: Leveraging Neural Radiance Fields for Tactile Sensory Data Generation". In: *6th Annual Conference on Robot Learning.* 2022. URL: `https://openreview.net/forum?id=No3mbanRlZJ`.

[31] Fengyu Yang, Jiacheng Zhang, and Andrew Owens. "Generating Visual Scenes from Touch". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).* Oct. 2023, pp. 22070–22080.

[32] Vedant Dave, Fotios Lygerakis, and Elmar Rueckert. "Multimodal Visual-Tactile Representation Learning through Self-Supervised Contrastive Pre-Training". In: *arXiv preprint arXiv:2401.12024* (2024).

[33] Francois R Hogan et al. "Tactile dexterity: Manipulation primitives with tactile feedback". In: *2020 IEEE international conference on robotics and automation (ICRA).* IEEE. 2020, pp. 8863–8869.

[34] Vedant Dave and Elmar Rueckert. "Predicting full-arm grasping motions from anticipated tactile responses". In: *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids).* IEEE. 2022, pp. 464–471.

[35] Yi Zheng et al. "Autonomous Learning of Page Flipping Movements via Tactile Feedback". In: *IEEE Transactions on Robotics* 38.5 (2022), pp. 2734–2749. DOI: `10.1109/TRO.2022.3168731`.

[36] Ziwei Xia et al. "A review on sensory perception for dexterous robotic manipulation". In: *International Journal of Advanced Robotic Systems* 19.2 (2022), p. 17298806221095974.

[37] Roberto Calandra et al. "The feeling of success: Does touch sensing help predict grasp outcomes?" In: *arXiv preprint arXiv:1710.05512* (2017).

[38] Roberto Calandra et al. "More Than a Feeling: Learning to Grasp and Regrasp Using Vision and Touch". In: *IEEE Robotics and Automation Letters* 3.4 (2018), pp. 3300–3307. DOI: `10.1109/LRA.2018.2852779`.

[39] Michelle A. Lee et al. "Making Sense of Vision and Touch: Learning Multimodal Representations for Contact-Rich Tasks". In: *IEEE Transactions on Robotics* 36.3 (2020), pp. 582–596. DOI: `10.1109/TRO.2019.2959445`.

[40] Stephen Tian et al. "Manipulation by Feel: Touch-Based Control with Deep Predictive Models". In: *2019 International Conference on Robotics and Automation (ICRA).* 2019, pp. 818–824. DOI: `10.1109/ICRA.2019.8794219`.

[41] Ting Chen et al. "A Simple Framework for Contrastive Learning of Visual Representations". In: *Proceedings of the 37th International Conference on Machine Learning.* Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 1597–1607.

[42] Kaiming He et al. "Momentum Contrast for Unsupervised Visual Representation Learning". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 2020, pp. 9726–9735. DOI: `10.1109/CVPR42600.2020.00975`.

[43] Tuomas Haarnoja et al. "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor". In: (2017).

[44] John Schulman et al. *Proximal Policy Optimization Algorithms.* 2017. arXiv: `1707.06347 [cs.LG]`.

[45] Aravind Srinivas, Michael Laskin, and Pieter Abbeel. "CURL: Contrastive Unsupervised Representations for Reinforcement Learning". In: *CoRR* abs/2004.04136 (2020). arXiv: `2004.04136`. URL: `https://arxiv.org/abs/2004.04136`.

[46] Michael Laskin et al. "Reinforcement Learming with Augmented Data". arXiv:2004.14990.

[47] Antonin Raffin et al. "Stable-Baselines3: Reliable Reinforcement Learning Implementations". In: *Journal of Machine Learning Research* 22.268 (2021), pp. 1–8. URL: `http://jmlr.org/papers/v22/20-1364.html`.