Multi-Modal Mutual Information (MuMMI) Training for Robust Self-Supervised Deep Reinforcement Learning

Kaiqi Chen*, Yong Lee*, and Harold Soh
Dept. of Computer Science, National University of Singapore.
{kaiqi, liy, harold}@comp.nus.edu.sg

***Abstract*— This work focuses on learning useful and robust deep world models using multiple, possibly unreliable, sensors. We find that current methods do not sufficiently encourage a shared representation between modalities; this can cause poor performance on downstream tasks and over-reliance on specific sensors. As a solution, we contribute a new multi-modal deep latent state-space model, trained using a mutual information lower-bound. The key innovation is a specially-designed density ratio estimator that encourages consistency between the latent codes of each modality. We tasked our method to learn policies (in a self-supervised manner) on multi-modal Natural MuJoCo benchmarks and a challenging Table Wiping task. Experiments show our method significantly outperforms state-of-the-art deep reinforcement learning methods, particularly in the presence of missing observations.**

I. Introduction

We live in a rich complex world. To make sense of it, humans (and other biological organisms) integrate information from a variety of senses. Our sensory apparatus (e.g., eyes, ears, skin) are often complementary, but also provide redundant information. This redundancy promotes robustness; biological agents display the incredible ability to cope under the temporary, or even permanent, loss of any given sense.

One might expect that artificial agents and robots can reap similar benefits from multiple sensory modalities. Indeed, many modern-day robots are equipped with a variety of sensors—e.g., cameras, microphones, tactile and proprioception sensors—that enable them to better perceive their environment. When combined with powerful representation learners (such as deep neural networks), these different sources of information can be used to learn world models for more robust decision-making and policy learning.

Unfortunately, learning robust world models from multiple raw sensory inputs remains challenging. Rather than improving performance, our preliminary deep reinforcement learning (RL) experiments revealed that including additional modalities can cause performance to *deteriorate*. The learned policies often failed to match the performance of a single-modality model, and were not robust to missing data.

In this work, we address the issue above and answer the question: *how can we learn complex world models from multiple, but possibly unreliable, sensors?* We develop a *modular* multi-modal deep latent state-space model (MSSM) that can be used for various robot tasks, including model-based RL and planning. Compared to deep models that "concatenate" different modalities [1], [2], structural modularity in our probabilistic graphical model provides a principled technique for dealing with missing data (rather than masking) with fewer parameters.

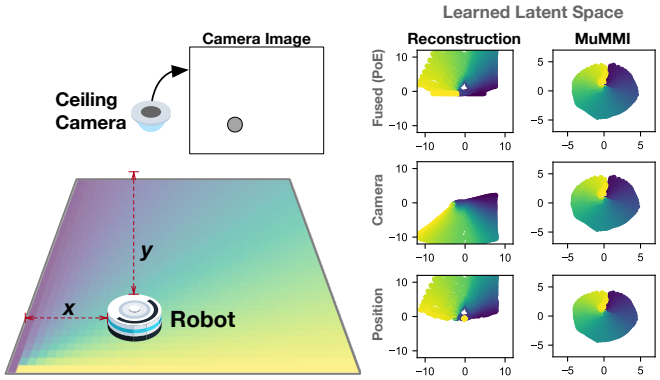

Fig. 1: A simple illustrative example of a robot in a 2D world with two sensory modalities: laser rangers that give its (x, y) position and a ceiling camera provides a scene image. A Deep Latent Space Model (SSM) trained using Product-of-Experts (PoE) fusion and a reconstruction-based loss did *not* learn a robust latent space from gathered data (left plots, colors indicate ground-truth position): the overlap between the two modality-specific latent spaces is small and model is over-reliant on the position sensor. The experts were "miscalibrated" in that the camera expert predicts a much higher variance relative to the position expert and thus, has little influence during PoE fusion. In contrast, our proposed MuMMI training encourages a consistent latent space across the different modalities (right plots) with calibrated experts.

Our key contribution is a mutual-information (MI) driven training method. Prior works have trained multimodal deep models by maximizing a reconstruction-based variational evidence lower-bound (ELBO) of the log data likelihood [3], [4], [5]. Our insight is that the standard ELBO does not sufficiently enforce a *shared* latent space between the different modalities. As a result, the learned world models do not well-integrate information from multiple sensors and the learned space is poorly structured (see Fig. 1 and additional plots in the online appendix [6]). As a remedy, we derive a MI-based lower-bound that is optimized via the InfoNCE loss [7]. Within this contrastive framework, we explicitly encourage the different modality networks to be consistent with one another via a specially-designed density ratio estimator. Unlike prior work on self-supervised RL with multiple

*Equal Contribution.

modalities [2], [8], our methodology is task-independent and alleviates the need to craft task/sensor-specific semi-supervised losses.

Experiments show that our Multi-Modal Mutual Information (MuMMI) approach significantly outperforms existing state-of-the-art techniques for self-supervised RL [5], [9] on Natural MuJoCo tasks [9] augmented with additional modalities. A further preliminary experiment on the challenging Robosuite Table Wiping task [10] shows that MuMMI is able to learn policies that are robust to a missing sensor. Specifically, inputs from two RGB cameras (one workspace camera and another mounted on the robot) were provided during training. During testing, we observed policy performance remained comparable even when completely removing the workspace camera.

In summary, this paper presents three key contributions:

- The Multi-Modal State-space model (MSSM), which can represent complex dynamics and multi-modal observations;
- The MuMMI training loss that encourages modalities to share a common latent space, which promotes robustness to missing observations;
- Empirical results showing that the MSSM trained with MuMMI outperforms competing methods and ablated variants, which indicate the importance of a modular structure and a shared latent space.

II. Preliminaries: Latent State-Space Models

Latent state-space models (SSMs) have been a long-standing staple of robotics. For example, the popular Kalman filter [11] comprises Gaussian latent (hidden) random variables with linear transitions between time-steps and linear observation functions. Other example SSMs include Hidden Markov Models [12] for discrete latent spaces, and probabilistic SLAM models [13]. This section assumes familiarity with probabilistic graphical models (PGMs); please refer to [14] for an excellent introduction.

Modern-day SSMs that leverage deep neural networks are able to capture complex nonlinear transitions and rich high-dimensional observations (e.g., camera images). Figure 2.A. illustrates a prototypical SSM where the z_t's are latent states from which the observations x_t's are generated. Transitions between time-steps t are Markovian and conditioned upon actions a_t taken by the robot. In reinforcement learning (RL) settings, we also include a reward per time-step r_t; here, we consider state-dependent reward distributions. Given the probabilistic graphical model in Fig. 2.A., the joint distribution of the model factorizes as:

$$p_\theta(x_{1:T}, r_{1:T}, z_{1:T}|a_{1:T}) = \prod_{t=1}^{T} p_\theta(x_t|z_t)p_\theta(r_t|z_t)p_\theta(z_t|z_{t-1}, a_{t-1}) \tag{1}$$

where θ are model parameters, $x_{1:T}$ denotes all observations from $t = 1, \ldots, T$, and likewise for $r_{1:T}$,$z_{1:T}$ and $a_{1:T}$. The three distributions in the factorization above correspond to:

$$\text{Observations:} \quad p_\theta(x_t|z_t) \tag{2}$$

$$\text{Rewards:} \quad p_\theta(r_t|z_t) \tag{3}$$

$$\text{Transitions:} \quad p_\theta(z_t|z_{t-1}, a_{t-1}) \tag{4}$$

and can be modelled using nonlinear function approximators such as deep neural networks.

One can view the model above as a Partially-Observable Markov Decision Process (POMDP) [15], [16] that is specified up to the unknown parameters θ. We would like to learn θ from observed data, $\mathcal{D} = \{x_t, a_t, r_t\}_{t=1}^{T}$, but maximum likelihood estimation is generally intractable as we need to marginalize out the latent z_t's. As such, we optimize the evidence lower bound (ELBO) under the data distribution p_d, i.e., $\mathbb{E}_{p_d}[\mathcal{L}_e] \leq \mathbb{E}_{p_d}[\log p_\theta(x_{1:T}, r_{1:T}|a_{1:T})]$, where

$$\mathcal{L}_e = \sum_{t=1}^{T} \Big(\mathop{\mathbb{E}}_{q_\phi(z_t)} [\log p_\theta(x_t|z_t)] + \mathop{\mathbb{E}}_{q_\phi(z_t)} [\log p_\theta(r_t|z_t)] - \mathop{\mathbb{E}}_{q_\phi(z_{t-1})} [\mathbb{D}_{\mathrm{KL}} [q_\phi(z_t) \| p_\theta(z_t|z_{t-1}, a_{t-1})]] \Big) \tag{5}$$

using a variational distribution q_ϕ, which is typically an *inference network* parameterized by ϕ. For simplicity, we denote the inference network as $q_\phi(z_t)$, but keep in mind the distribution is often conditioned on observations, e.g., $q_\phi(z_t|x_t)$. In the ELBO, the first two reconstruction terms encourage encoding of information of x_t and r_t in the latent state z_t. The third Kullback-Leibler (KL) divergence term enforces consistency between the variational distribution q_ϕ and the transition dynamics $p_\theta(z_t|z_{t-1}, a_{t-1})$.

III. Multi-Modal Deep Latent State-Space Model

In this section, we describe our Multi-modal state-space model (MSSM), which extends the aforementioned SSM to multiple sensory modalities. We first describe the model structure, and then proceed to detail our MI-based training methodology.

Model Structure. As a guide, Fig. 2.B. illustrates a two-time-slice view of our model. Compared to the vanilla variant in Fig. 2.A., MSSM generates multiple observations corresponding to the M different modalities (x_t^m in the plates) and employs a modified Recurrent SSM (RSSM) structure [17] — we decompose the latent state z_t into three variables $z_t = \left[h_t, s_t^c, s_t^f\right]$. This splits the latent state into deterministic and stochastic parts; the transition governing h_t is deterministic, which helps the model better remember previous states. Unlike prior work [17], we further decompose the stochastic variable: s_t^f encodes information about the current observations across modalities, whilst the "combined" stochastic variable s_t^c also encodes past information. We find that this decomposition, when combined with appropriate inference networks, enables faster and more stable training. The joint distribution of the model factorizes

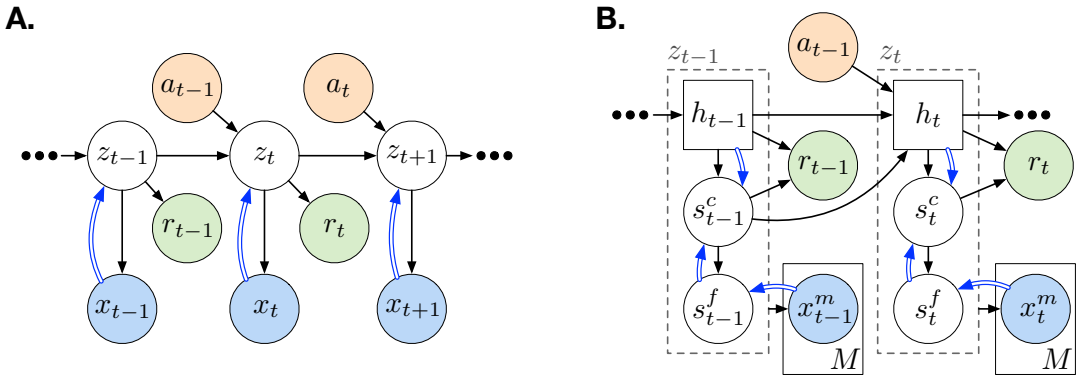

Fig. 2: Probabilistic Graphical Models (PGMs) for (**A.**) the basic latent state-space model (SSM) used in reinforcement learning/planning contexts and (**B.**) our Multi-Modal State-Space Model (MSSM). Circle nodes represent random variables and shaded nodes are observed. The square nodes indicate a deterministic function mapping. The MSSM generalizes the basic model to M modalities and decomposes the latent variables $z_t = [h_t, s_t^c, s_t^f]$. Inference networks q_ϕ are shown using blue double-lined arrows; we see that $q(s_t^f|x_t^{1:M})$ fuses information across the modalities, whilst $q(s_t^c|s_t^f, h_t)$ encodes information from the latent dynamics and the fused observations. Please see main text in Secs. II and III for additional details.

in a similar manner to eq. (1):

$$p_\theta(x_{1:T}^{1:M}, r_{1:T}, z_{1:T}|a_{1:T}) = \prod_{t=1}^{T}\prod_{m=1}^{M} p_\theta(x_t^m|z_t)p_\theta(r_t|z_t)p_\theta(z_t|z_{t-1}, a_{t-1}) \quad (6)$$

where $x_{1:T}^{1:M}$ denotes all the M observations at every time step $t = 1, \ldots, T$, and

$$\text{Observations:} \quad p_\theta(x_t^m|z_t) = p_\theta(x_t^m|s_t^c, h_t) \quad (7)$$

$$\text{Rewards:} \quad p_\theta(r_t|z_t) = p_\theta(r_t|s_t^c, h_t) \quad (8)$$

$$\text{Transitions:} \quad p_\theta(z_t|z_{t-1}, a_{t-1}) = p_\theta(s_t^f|s_t^c) \, p_\theta(s_t^c|h_t) g_\theta(h_t|s_{t-1}^c, h_{t-1}, a_{t-1}) \quad (9)$$

Note that the function g above is deterministic (indicated by squares in Fig. 2.B.).

Model Training via Standard ELBO. As in the single modality case, a possible training option is to maximize the ELBO,

$$\log p_\theta(x_{1:T}^{1:M}, r_{1:T}|a_{1:T}) \geq \mathcal{L}_e^M$$
$$= \sum_{t=1}^{T}\Big(\sum_{m=1}^{M}\mathbb{E}_{q_\phi(z_t)}\left[\log p_\theta(x_t^m|z_t)\right] + \mathbb{E}_{q_\phi(z_t)}\left[\log p_\theta(r_t|z_t)\right] - \mathbb{E}_{q_\phi(z_{t-1})}\left[\mathbb{D}_{\mathrm{KL}}\left[q_\phi(z_t)\|p_\theta(z_t|z_{t-1}, a_{t-1})\right]\right]\Big) \quad (10)$$

using the variational distribution,

$$q_\phi(z_{1:T}|x_{1:T}^{1:M}, a_{1:T}) = \prod_{t=1}^{T} q(z_t|z_{t-1}, x_t^{1:M}, a_{t-1})$$
$$= \prod_{t=1}^{T}\left[\prod_{m=1}^{M} q(s_t^f|x_t^m)\right] q(s_t^c|s_t^f, h_t) g_\theta(h_t|s_{t-1}^c, h_{t-1}, a_{t-1}) \quad (11)$$

where $q(s_t^f|x_t^m)$, $q(s_t^c|s_t^f, h_t)$ and $p(s_t^c|h_t)$ are Gaussians, and the different modalities are fused via a Product-of-Experts (PoE) [18]. If modality m is missing, we can simply drop corresponding expert $q(s_t^f|x_t^m)$.

One key problem with maximizing the ELBO above is that the objective is under-constrained: the different modality experts need not share the same latent space. Prior work has primarily resorted to randomly dropping modalities during training to force consistency, but our experiments showed this approach may not be robust (Fig. 1).

Model Training via MuMMI. In this work, we pursue an alternative information-theoretic approach, which turns out to be equivalent to maximizing $\mathcal{L}_e^M$ under specific assumptions. Let us define $v_{1:T}^{\setminus m} = (x_{1:T}^{\setminus m}, z_{1:T})$ where $x_{1:T}^{\setminus m}$ denotes observations from all the modalities *except* modality m. To reduce clutter, we will drop the explicit dependence on $a_{1:T}$. Assume that the data is generated from the MSSM and consider the mutual information between $x_{1:T}^m$ and $v_{1:T}^{\setminus m}$:

$$\mathbb{I}[x_{1:T}^m; v_{1:T}^{\setminus m}] = \sum p(x_{1:T}^m, v_{1:T}^{\setminus m}) \log \frac{p(x_{1:T}^m, v_{1:T}^{\setminus m})}{p(x_{1:T}^m)p(v_{1:T}^{\setminus m})}$$
$$= \sum p(x_{1:T}^{1:M})p(z_{1:T}|x_{1:T}^{1:M}) \log \frac{p(x_{1:T}^m|z_{1:T})}{p(x_{1:T}^m)}$$
$$= \mathbb{E}_{p(x_{1:T}^{1:M})p(z_{1:T}|x_{1:T}^{1:M})}\left[\log p(x_{1:T}^m|z_{1:T})\right] - C^m \quad (12)$$

where $C^m = \mathbb{E}_{p(x_{1:T}^m)}\left[p(x_{1:T}^m)\right]$, and we have leveraged the conditional independence assumptions in the MSSM when dropping the dependence on $x_{1:T}^{\setminus m}$ in $p(x_{1:T}^m|x_{1:T}^{\setminus m}, z_{1:T})$. Intuitively, $\mathbb{I}[x_{1:T}^m; v_{1:T}^{\setminus m}]$ captures the mutual dependence between a given modality m and the remaining observations together with the latent state. If we assume that $q(z_{1:T}|x_{1:T}^m) = p(z_{1:T}|x_{1:T}^{1:M})$, we can combine eq. (10) and eq. (12) to yield

$$\mathbb{E}_{p_d}[\mathcal{L}_e^M] = \sum_{m=1}^{M}\left(\mathbb{I}[x_{1:T}^m; v_{1:T}^{\setminus m}] + C^m\right) + \mathbb{E}_{p_d q_\phi(z_t)}\left[\log p_\theta(r_t|z_t)\right] - \mathbb{E}_{p_d q_\phi(z_{t-1})}\left[\mathbb{D}_{\mathrm{KL}}\left[q_\phi(z_t)\|p_\theta(z_t|z_{t-1}, a_{t-1})\right]\right] \quad (13)$$

which relates $\mathbb{I}[x_{1:T}^m; v_{1:T}^{\setminus m}]$ to the ELBO. For the purposes of learning, $\sum_m C^m$ is a constant that does not depend on the parameters θ, and can be dropped.

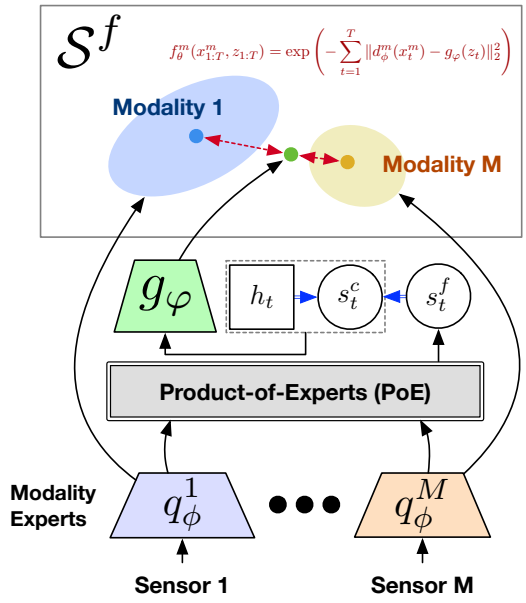

Fig. 3: MuMMI training uses a density ratio estimator f_θ^m (eq. (14)) that acts to minimize the squared distances between the mean of each modality expert and a transformed fused latent code. This encourages the experts to project to points in a shared latent space.

To optimize $\mathbb{I}[x_{1:T}^m; v_{1:T}^{\setminus m}]$, we use the InfoNCE loss [7]. Let us define the density ratio estimator,

$$f_\theta^m(x_{1:T}^m, v_{1:T}^{\setminus m}) \propto \frac{p(x_{1:T}^m | v_{1:T}^{\setminus m})}{p(x_{1:T}^m)} = \frac{p(x_{1:T}^m | z_{1:T})}{p(x_{1:T}^m)} \tag{14}$$

where we have again exploited the conditional independence between modalities given $z_{1:T}$. As such, we can specify M density ratio estimators independently for each modality, which simplifies our setup and eases computational burden. Abusing notation, we let $f_\theta^m(x_{1:T}^m, z_{1:T}) = f_\theta^m(x_{1:T}^m, v_{1:T}^{\setminus m})$. From [7], we can show that,

$$\mathbb{I}[x_{1:T}^m; v_{1:T}^{\setminus m}] \geq \mathbb{E}\left[\log \frac{f_\theta^m(x_{1:T}^{+,m}, z_{1:T})}{\sum_{x_{1:T}^{-,m}} f_\theta^m(x_{1:T}^{-,m}, z_{1:T})}\right] \tag{15}$$

where $x_{1:T}^{+,m}$ and $x_{1:T}^{-,m}$ are "positive" and "negative" samples, respectively. We obtain the positive sample by drawing $x_{1:T}^{+,m} \sim p_\theta(x_{1:T}^m | z_{1:T})$ and $N-1$ negative samples from the proposal distribution $p_d(x_{1:T}^m)$.

Although the InfoNCE is a looser bound of the log marginal likelihood, it affords us additional design freedom in the density ratio estimator f_θ^m. We propose a design that encourages *each* modality expert to map data to points close to a (fused) latent state:

$$f_\theta^m(x_{1:T}^m, z_{1:T}) = \exp\left(-\sum_{t=1}^{T} \|d_\phi^m(x_t^m) - g_\varphi(z_t)\|_2^2\right) \tag{16}$$

where d_ϕ^m and g_φ are also neural networks. We set d_ϕ^m to share parameters with $q_\phi(s_t^f | x_t^m) = \mathcal{N}(\mu_\phi^m(x_t^m), v_\phi^m(x_t^m))$, i.e., $d_\phi^m = \mu_\phi^m$. Here, f can be seen as a squared exponential kernel and maximizing the numerator in eq. (15) across modalities encourages consistent projections (see Fig. 3).

Final Loss and In-Practice. Training the MSSM via MuMMI entails optimizing $\mathbb{E}_{p_d}[\widehat{\mathcal{L}}]$ where:

$$\begin{aligned}\mathbb{E}_{p_d}[\widehat{\mathcal{L}}] = \sum_{m=1}^{M} \lambda^m \mathbb{E}\left[\log \frac{f_\theta^m(x_{1:T}^{+,m}, z_{1:T})}{\sum_{x_{1:T}^{-,m}} f_\theta^m(x_{1:T}^{-,m}, z_{1:T})}\right] \\ + \mathbb{E}_{p_d q_\phi(z_t)}\left[\log p_\theta(r_t | z_t)\right] \\ - \mathbb{E}_{p_d q_\phi(z_{t-1})}\left[\mathbb{D}_{\mathrm{KL}}\left[q_\phi(z_t) \| p_\theta(z_t | z_{t-1}, a_{t-1})\right]\right]\end{aligned} \tag{17}$$

and the λ^m's are hyperparameters, which can be set equal or tuned using prior knowledge of which modality is more informative. To compute f_θ^m, we use a strategy similar to prior work [9]: we sample a batch of sequences $\{x_{1:T}^{1:M,i}, a_{1:T}^i, r_{1:T}^i\}_{i=1}^B$ from a replay buffer, where B is the batch size. For each state-observation pair, we treat the other $(B \times T) - 1$ observations in the same batch as negative samples.

IV. Related Work

Our work builds upon recent advances in probabilistic multi-modal models and deep reinforcement learning. Specifically, MuMMI uses PoE fusion [18], which was previously used in a multi-modal variational autoencoder [3] that was later extended to sequential settings [4]. Multi-modal models have also been adopted in robotics applications, where feature vectors from different modalities are concatenated into a single latent representation [1], [2]. Lately, PoE-based fusion has been applied to multi-modal self-supervised training [8], but unlike MuMMI, the method relies on hand-crafted task-dependent losses. In a related research thread, very recent work has explored event-driven multi-modal representations using Spiking Neural Networks [19]. Here, we use deep artificial neural networks but MuMMI can potentially be extended to event-driven learning.

MuMMI is also related to recent self-supervised model-based RL methods, e.g., PlaNET [17] and Dreamer [5], which learn latent dynamics models via interactions with the environment. The backbone of these methods is the RSSM model, on which our MSSM is based. However, these techniques rely the standard reconstruction-based ELBO, which is not robust to irrelevant noise. Our approach is closely related to the recently proposed CVRL [9], which learns using the InfoNCE loss. However, CVRL (and other self-supervised RL methods) have largely focused on single-modality learning with reliable sensors. Unlike the works above, MuMMI trains a multi-modal world model (the MSSM) that is demonstrably robust to missing data.

V. Experiment: Multi-Modal Natural MuJoCo

In this section, we describe experiments designed to evaluate the MSSM and MuMMI on the task of self-supervised RL; for simplicity, we will refer to the MSSM with MuMMI training as MuMMI. Our goal was to ascertain whether MuMMI led to better performance and robustness to missing data, compared to competing state-of-the-art methods.

TABLE I: Multi-Modal Natural Mujoco Experiment. Performance measured by Mean Total Episode Reward (averaged over 30 episodes, with standard deviations). Highest average reward in **bold**.

Task	Missing Data	MuMMI	MuMMI-b	CVRL	Dreamer
walker run	None	**466.7** $\pm$ 25.5	116.4 $\pm$ 13.6	311.7 $\pm$ 30.2	71.3 $\pm$ 12.4
	Medium	**438.7** $\pm$ 33.1	119.1 $\pm$ 12.1	277.5 $\pm$ 22.4	70.7 $\pm$ 9.70
	High	**382.4** $\pm$ 38.5	105.2 $\pm$ 14.6	255.8 $\pm$ 31.4	70.9 $\pm$ 37.3
walker walk	None	**954.5** $\pm$ 21.5	870.4 $\pm$ 44.1	732.7 $\pm$ 61.7	182.9 $\pm$ 37.7
	Medium	**939.9** $\pm$ 26.1	813.1 $\pm$ 65.1	703.4 $\pm$ 48.9	163.1 $\pm$ 31.8
	High	**890.0** $\pm$ 50.1	760.3 $\pm$ 61.5	607.5 $\pm$ 58.3	172.8 $\pm$ 39.2
walker stand	None	**966.8** $\pm$ 24.6	959.8 $\pm$ 34.5	955.6 $\pm$ 27.6	307.0 $\pm$ 54.40
	Medium	**952.4** $\pm$ 28.7	946.5 $\pm$ 24.1	940.1 $\pm$ 34.3	334.3 $\pm$ 122.6
	High	**922.0** $\pm$ 59.7	918.2 $\pm$ 58.4	918.2 $\pm$ 31.9	280.4 $\pm$ 72.70
finger spin	None	**965.6** $\pm$ 10.9	679.1 $\pm$ 38.7	705.9 $\pm$ 32.1	342.7 $\pm$ 37.3
	Medium	**956.7** $\pm$ 17.0	661.4 $\pm$ 42.7	685.2 $\pm$ 36.4	303.6 $\pm$ 30.7
	High	**929.1** $\pm$ 20.8	620.4 $\pm$ 58.2	623.8 $\pm$ 39.1	271.2 $\pm$ 35.1
cup catch	None	**949.5** $\pm$ 34.3	725.8 $\pm$ 249.2	922.0 $\pm$ 48.1	124.0 $\pm$ 290.3
	Medium	**948.0** $\pm$ 32.9	637.8 $\pm$ 335.3	904.2 $\pm$ 78.7	47.7 $\pm$ 147.3
	High	**927.3** $\pm$ 38.0	540.5 $\pm$ 317.7	922.3 $\pm$ 47.9	73.5 $\pm$ 225.3

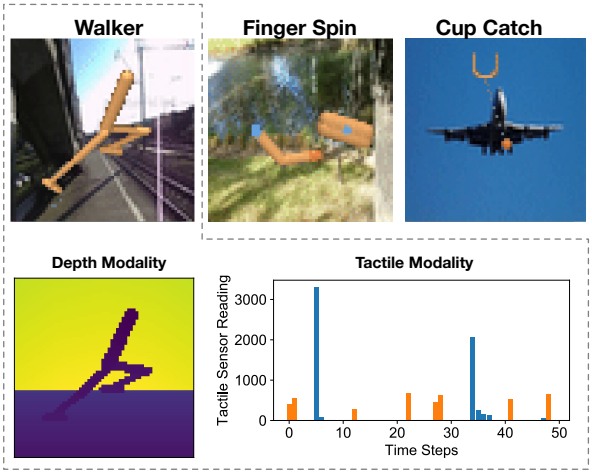

Fig. 4: Natural Mujuco environments and modalities used in our experiments (**top**) RGB images for walker stand/walk/run, finger spin and cup catch. The background images are continuously changing. (**bottom**) Additional two modalities for the walker, i.e., depth image and tactile sensor. The tactile sensors were placed on the soles of the feet and activated when it came into contact with the ground. For the other two environments, the tactile sensor was placed on the finger-tip and inside the cup, respectively.

Methods. We compare MuMMI against two representative state-of-the-art model-based deep RL methods: Dreamer [5] and CVRL [9]. Dreamer uses a reconstruction-based ELBO, whilst CVRL is trained using a contrastive loss (but without a product-of-experts fusion layer). For both models, feature vectors extracted from modality-specific deep networks are fused via concatenation and missing observations are masked with zeros (similar to prior work [4], [1]). We also tested MuMMI-b; a variant of MuMMI with a modified density ratio estimator: $f_b(x_{1:T}^{1:M}, z_{1:T}) = \exp\left(-\sum_{t=1}^{T} \|b_\phi(x_t^{1:M}) - g_\varphi(z_t)\|_2^2\right)$, where $b_\phi(x_t^{1:M})$ is set to the mean of the fused PoE distribution. Compared to eq. (16), f_b promotes consistency between the PoE-fused latent vectors and the learned dynamics. It does not directly constrain *individual* modalities have similar latent codes, but may work well if given sufficient data (and trained using random drops [3]). All methods used latent imagination, an actor-critic RL method [5] and latent-guided MPC [9].

Multi-Modal Tasks. We used the MuJoCo-powered DeepMind Control Suite [20], but augmented to have complex backgrounds (Natural MuJoCo [9]) and additional modalities (Fig. 4). The standard benchmarks already pose challenges common to robot learning: sparse rewards, high-dimensional 3D scenes, many degrees of freedom, and contact dynamics. The complex backgrounds—videos from ILSVRC dataset [21]—add a degree of realism and difficulty as the robot needs to separate useful information from irrelevant noise. We selected 5 benchmark tasks based on available computational budget. The modalities for all tasks comprise RGB and depth images, and tactile feedback. The backgrounds are assumed far and do not appear in the depth images; this tests if the models are able to use this "clean" modality to improve performance, yet not become overly reliant on it. The tactile modality has significantly different properties compared to the images; it is a sparse signal that occurs when certain parts (i.e., the walker feet, finger tip, and inside-cup) come into contact with the ground or other objects.

Methodology. For each task-method pair, we conducted 3 training sessions where each session was initialized with a different random seed and trained for 2 million episodes. Each session took $\approx$ 1 day to complete on a workstation with a Nvidia 2080Ti GPU. During training, data was randomly dropped to simulate data loss (e.g., from faulty sensors or occlusions); for each modality, we dropped segments of varying lengths (the start and length of missing segments are uniformly random, but constrained so that the missing rate was 37.5% of the complete data). In the testing stage, we compared each method's accumulated rewards per-episode (averaged over the 3 trained policies). Each policy was tested over 3 batches of 10 episodes, where each batch with a different missing rate (None: 0%; Medium: 37.5%; High: 75%). Complete model architecture details and source code is available in the online appendix.

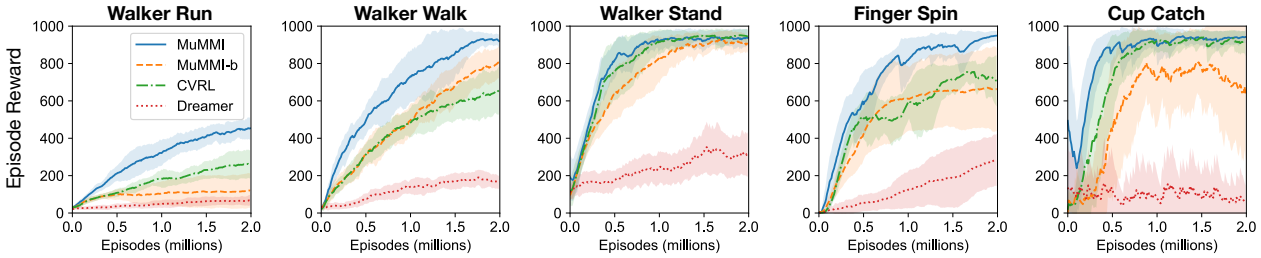

Fig. 5: Model performance across training episodes (moving average smoothing with weight 0.9). The performance curves are steeper for MuMMI on a majority of the tasks, indicating faster learning compared to MuMMI-b and competing approaches (Dreamer and CVRL).

TABLE II: Table Wiping Performance measured by Mean Total Episodic Reward (averaged over 30 episodes, with standard deviations)

Modalities	MuMMI	MuMMI-b	MSSM-e
All Modalities (Medium Missing Data)	61.4 ± 70.4	48.8 ± 40.5	60.9 ± 56.1
All Modalities (Full Observed)	60.6 ± 69.9	47.9 ± 37.2	64.9 ± 85.3
Robot Camera (Medium Missing Data)	57.8 ± 69.8	62.1 ± 93.7	64.0 ± 95.9
Robot Camera (Full Observed)	65.9 ± 76.9	69.2 ± 97.2	60.4 ± 98.2

Results. The final performance of the different models is summarized in Table I. On all of the tasks, MuMMI outperforms all other competing approaches by a significant margin. The poorer performance of the ablated MuMMI-ab indicates the importance of a common latent space for PoE fusion. We observed that MuMMI degrades gracefully with greater amount of missing data, but remains robust compared to the other methods. Between the concatenation fusion methods (Dreamer and CVRL), Dreamer has poorer performance, despite given access to the clean depth images[1]. In comparison, CVRL was better able to learn from multiple modalities; we posit that Dreamer reconstruction loss does not permit the model to neglect the irrelevant inputs, which hampered learning of a good latent code. Finally, we observed that MuMMI learns faster than other methods, as indicated by the steeper learning curves in Fig. 5.

VI. Case Study: Table Wiping

In this section, we describe preliminary experiments using MuMMI to train a Franka-Emika Panda arm on the challenging Table Wiping benchmark task [10]. Due to space constraints, we describe the essentials; please see the online appendix for additional information. We compared three methods: MuMMI, MuMMI-b and MSSM-e. MSSM-e is trained using a reconstruction-based ELBO (similar to Dreamer), but uses PoE instead of concatenation to fuse the modalities. We trained MSSM-e using a similar approach as [3] where missing input modalities are dropped.

In the Table Wiping task, the Panda robot has to clean a table by erasing markings on its surface. The markings are randomized at the start of each episode. This task is one of the more challenging benchmarks in Robosuite and previous work using the state-of-the-art model-free soft-actor critic (SAC) [22] failed solve the task [10]. Here, the robot can access two modalities: a RGB camera mounted on the top of the robot and a workspace RGB camera (Fig. 6).

[1]Given a single modality of clean image data, Dreamer is generally able to achieve high rewards on the tasks tested [5], [9].

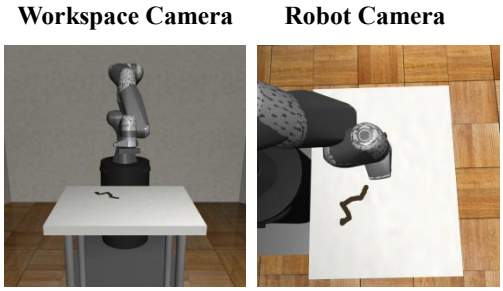

Fig. 6: Views from the two RGB cameras used for Table Wiping.

We trained each method for 1 million episodes with domain randomization and moderate data loss (37.5%) during training. In the testing stage, we compared each method's accumulated rewards per-episode (averaged over 30 episodes). Our results are summarized in Table II. We see that the methods were robust to removal of the workspace camera; performance was not drastically affected by the removal. Interestingly, we see that MSSM-e was also able to perform well for this particular problem. These preliminary results are promising; they show MuMMI and MSSM can be applied towards robotics problems in scenarios with unreliable sensors.

VII. Conclusions

This work presents the MSSM and MuMMI. Together, they can be used to learn robust world-models from multimodal sensory streams, even with significant amounts of missing data. Moving forward, we plan to apply MuMMI beyond self-supervised RL to other robot tasks including planning, human modeling, and imitation learning.

Acknowledgements

This work was supported by the Science and Engineering Research Council, Agency of Science, Technology and Research, Singapore, through the National Robotics Program under Grant No. 192 25 00054.

REFERENCES

[1] M. Zambelli, A. Cully, and Y. Demiris, "Multimodal representation models for prediction and control from partial information," *Robotics and Autonomous Systems*, vol. 123, p. 103312, 2020.

[2] 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.

[3] M. Wu and N. Goodman, "Multimodal generative models for scalable weakly-supervised learning," in *Advances in Neural Information Processing Systems*, 2018, pp. 5575–5585.

[4] T. Zhi-Xuan, H. Soh, and D. C. Ong, "Factorized inference in deep markov models for incomplete multimodal time series," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[5] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, "Dream to control: Learning behaviors by latent imagination," *arXiv preprint arXiv:1912.01603*, 2019.

[6] K. Chen, Y. Lee, and H. Soh, "Multi-modal mutual information (mummi) training for robust self-supervised deep reinforcement learning: Online appendix," 2021. [Online]. Available: https://clear-nus.github.io/project/mummi

[7] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[8] M. A. Lee, Y. Zhu, P. Zachares, M. Tan, K. Srinivasan, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Learning multimodal representations for contact-rich tasks," *IEEE Transactions on Robotics*, 2020.

[9] X. Ma, S. Chen, D. Hsu, and W. S. Lee, "Contrastive variational model-based reinforcement learning for complex observations," *arXiv preprint arXiv:2008.02430*, 2020.

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

[11] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Fluids Engineering*, 1960.

[12] L. Rabiner and B. Juang, "An introduction to hidden markov models," *ieee assp magazine*, vol. 3, no. 1, pp. 4–16, 1986.

[13] S. Thrun, W. Burgard, and D. Fox, "Probabilistic robotics." 2005.

[14] M. I. Jordan, "An introduction to probabilistic graphical models," *University of California, Berkeley*, 2003.

[15] K. J. Astrom, "Optimal control of markov processes with incomplete state information," *Journal of mathematical analysis and applications*, vol. 10, no. 1, pp. 174–205, 1965.

[16] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial intelligence*, vol. 101, no. 1-2, pp. 99–134, 1998.

[17] D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson, "Learning latent dynamics for planning from pixels," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2555–2565.

[18] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[19] T. Taunyazoz, W. Sng, H. H. See, B. Lim, J. Kuan, A. F. Ansari, B. Tee, and H. Soh, "Event-driven visual-tactile sensing and learning for robots," in *Proceedings of Robotics: Science and Systems*, July 2020.

[20] Y. Tassa, S. Tunyasuvunakool, A. Muldal, Y. Doron, S. Liu, S. Bohez, J. Merel, T. Erez, T. Lillicrap, and N. Heess, "dm_control: Software and tasks for continuous control," 2020.

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[22] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, *et al.*, "Soft actor-critic algorithms and applications," *arXiv preprint arXiv:1812.05905*, 2018.

APPENDIX

A. Latent Imagination and Actor-Critic

Following [5], after training, the agent generates the imagined trajectories using the learnt world model. Specifically, given a current state z_t, the agent sample the next imagined state by $\tilde{z}_{t+1} \sim p_\theta(z_{t+1}|z_{t-1}, a_{t-1})$ and associated reward $\tilde{r}_{t+1} \sim p_\theta(r_{t+1}|z_{t+1})$ and the next action $\tilde{a}_t \sim \pi_\eta(a_{t+1}|z_{t+1})$. This process is repeated until an imagined trajectory $\{\tilde{z}_t, \tilde{a}_t, \tilde{r}_t\}_{t=\tau}^{\tau+H}$ is generated. Then, the agent learns the action and value models by optimizing:

$$\begin{aligned} &\max_{\eta} \mathop{\mathbb{E}}_{\pi_\eta, p_\theta} \left(\sum_{t=\tau}^{\tau+H} V_\lambda(z_t) \right) \\ &\min_{\psi} \mathop{\mathbb{E}}_{\pi_\eta, p_\theta} \left(\sum_{t=\tau}^{\tau+H} \frac{1}{2} \left\| v_\psi(z_t) - V_\lambda(z_t) \right\|^2 \right) \end{aligned} \tag{18}$$

where

$$\begin{aligned} V_\lambda(\tilde{z}_t) &= \mathop{\mathbb{E}}_{\pi_\eta, p_\theta} \left(\sum_{n=\tau}^{h-1} \lambda_{n-\tau} \tilde{r}_n \right) \\ V_\lambda(\tilde{z}_t) &= (1-\lambda) \sum_{n=1}^{H-1} \lambda^{n-1} V_N^n(\tilde{z}_t) + \lambda^{H-1} V_N^H(\tilde{z}_t) \end{aligned} \tag{19}$$

The objective $\max_\eta \mathop{\mathbb{E}}_{\pi_\eta, p_\theta} \left(\sum_{t=\tau}^{\tau+H} V_\lambda(z_t) \right)$ optimizes the policy π_η under current critic and the objective $\min_\psi \mathop{\mathbb{E}}_{\pi_\eta, p_\theta} \left(\sum_{t=\tau}^{\tau+H} \frac{1}{2} \left\| v_\psi(z_t) - V_\lambda(z_t) \right\|^2 \right)$ optimizes the value estimation. We also use latent guided model predictive control as in [9].

All in all, in each iteration, MuMMI first learns the world model by using samples in replay buffers. Then, MuMMI use latent imagination to optimize the actor and critic. By iterating this process, the agent is able to learn behaviors in complex environments. To assist policy optimization, latent-guided MPC [9] was also used.

B. Multi-Modal State-Space Model

We use the similar model architectures similar to [5].

1) Transition Network: We use a GRU module to model the deterministic transition function g_θ. The dimension of h_t is 200 in both Multi-Modal Natural Mujoco and for Table Wiping tasks. We use a multi-layer perceptrons to model $p_\theta(s_t^c|h_t)$. $p_\theta(s_t^c|h_t)$ is modeled as a Gaussian with a diagonal covariance matrix. The multi-layer perceptron takes in h_t as an input and outputs the mean and variance of the $p_\theta(s_t^c|h_t)$ (Fig.7).

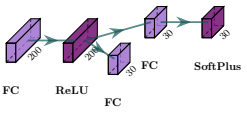

Fig. 7: The network modeling $p_\theta(s_t^c|h_t)$.

2) Inference Networks: We model $q(s_t^f|x_t^m)$ as Gaussian distributions with a diagonal covariance matrix. For each modality, we first use a network to extract features from raw data (Fig. 8 or Fig. 9) and then use another network to map this features to the mean and variance of $q(s_t^f|x_t^m)$ (Fig. 10). The dimension of s_t^f is 1024 in Multi-Modal Natural Mujoco tasks and 256 in the Table Wiping task. $q(s_t^c|s_t^f, h_t)$ is modeled as multi-layer perceptron, which takes in the concatenation of s_t^f, h_t as an input and outputs the mean and variance of $q(s_t^c|s_t^f, h_t)$ (Fig.11). The dimension of s_t^c is 30 in both Multi-Modal Natural Mujoco and for Table Wiping tasks.

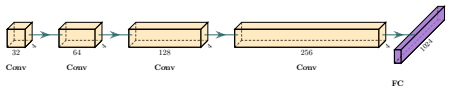

Fig. 8: Embedding networks for RGB image (or depth image)

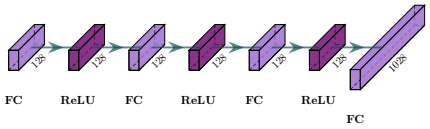

Fig. 9: Embedding networks for tactile

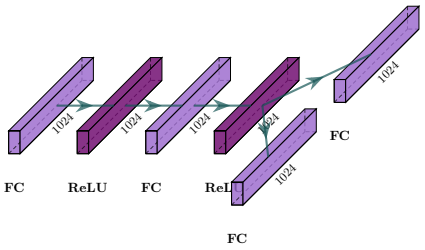

Fig. 10: Network that maps extracted features to mean and variance of $q(s_t^f|x_t^m)$

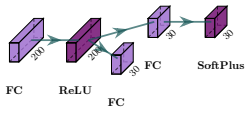

Fig. 11: Network that takes in the concatenation of s_t^f, h_t and outputs the mean and variance of $q(s_t^c|s_t^f, h_t)$

3) Actor and Value Networks: We used a multi-layer perceptron to model actor $\pi(a|z)$ and value function. We modelled the actor $\pi(a|z)$ as a Gaussian distribution with a diagonal covariance matrix. The actor networks takes latent states z as input and outputs the mean and variance for $\pi(a|z)$ (Fig.12). Similarly, the value network takes latent states z as input and outputs the values for value function (Fig.13).

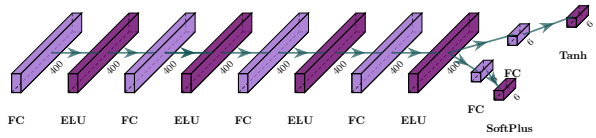

Fig. 12: Actor network that models $\pi(a|z)$.

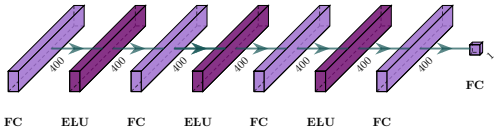

Fig. 13: Value network that models value function $V_\lambda(z)$.

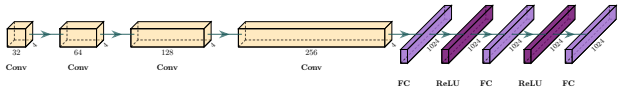

Fig. 14: Networks that extract features from RGB image or depth image.

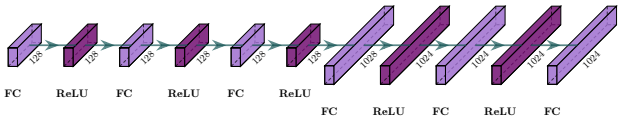

Fig. 15: Networks that extract features from tactile.

C. Baseline Models

Similar network structures were used in the baseline models. However, instead the product-of-experts (PoE), we use the networks in Fig. 14 and Fig. 15 to first extract features from different modalities, which are contacted and fed into another network (Fig.11) to obtain the mean and variance of $q(s_t^c|s_t^f, h_t)$. The dimension of h_t and s_t^f are the same as in MuMMI. In Multi-Modal Natural Mujoco tasks, the s_t^f is of dimension $3072 = 1024 \times 3$ (the concatenation of feature vectors extracted from three modalities—RGB image, depth image and tactile). In the Table Wiping task, the s_t^f is of dimension $512 = 256 \times 2$, which is the concatenation of feature vectors extracted from two modalities (RGB image from workspace camera and RGB image from robot camera).

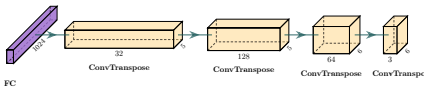

Fig. 16: Decoder network for RGB image.

For Dreamer [5], we use decoder networks to model $p_\theta(x_t^m|z_t)$, which is assumed to be Gaussian with diagonal covariance in our experiments. The decoder networks takes in latent state as an input and outputs the mean $p_\theta(x_t^m|z_t)$ (Fig. 16, Fig. 17 and Fig. 18). The variance of $p_\theta(x_t^m|z_t)$ is set as 1.0 for all modalities.

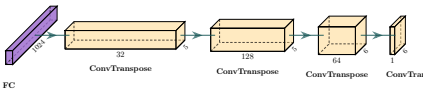

Fig. 17: Decoder network for depth image.

D. Additional Results

Additional results for the toy example can be seen in Fig. 19 and Fig. 20. In Fig. 19, we can see that using MuMMI results in a consistent representation among two camera but using the reconstruction loss does not. Also, if the two modalities are independent (the x and y positions of the robot), MuMMI can still learn a structured latent space. (Fig.20).

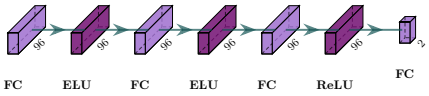

Fig. 18: Decoder network for tactile with 2 channel.

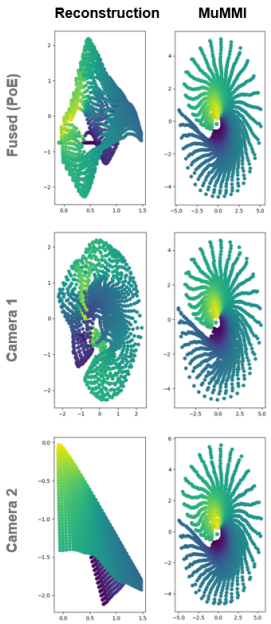

Fig. 19: The robot in a 2D world with two sensory modalities: a camera mounted in the front and a camera mounted behind. The figures show the latent space learnt by reconstruction and MuMMI.

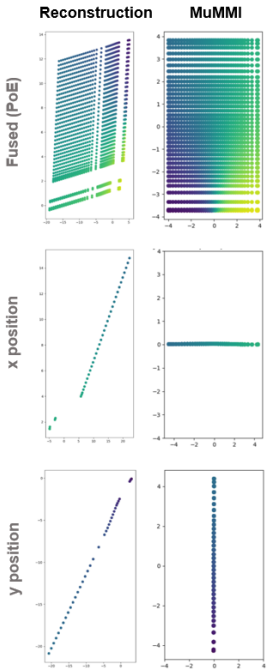

Fig. 20: The robot in a 2D world with two independent sensory modalities: x position of the robot and y position. The figures show the latent space learnt by reconstruction and MuMMI.