Bayesian Imitation Learning for End-to-End Mobile Manipulation

Yuqing Du [1] Daniel Ho [2] Alexander A. Alemi [3] Eric Jang [4] Mohi Khansari [2]

Abstract

In this work we investigate and demonstrate benefits of a Bayesian approach to imitation learning from multiple sensor inputs, as applied to the task of opening office doors with a mobile manipulator. Augmenting policies with additional sensor inputs—such as RGB + depth cameras—is a straightforward approach to improvin grobot perception capabilities, especially for tasks that may favor different sensors in different situations. As we scale multi-sensor robotic learning to unstructured real-world settings (e.g. offices, homes) and more complex robot behaviors, we also increase reliance on simulators for cost, efficiency, and safety. Consequently, the sim-to-real gap across multiple sensor modalities also increases, making simulated validation more difficult. We show that using the Variational Information Bottleneck (Alemi et al., 2016) to regularize convolutional neural networks improves generalization to held-out domains and reduces the sim-to-real gap in a sensor-agnostic manner. As a side effect, the learned embeddings also provide useful estimates of model uncertainty for each sensor. We demonstrate that our method is able to help close the sim-to-real gap and successfully fuse RGB and depth modalities based on an understanding of the situational uncertainty of each sensor. In a real-world office environment, we achieve 96% task success, improving upon the baseline by +16%.

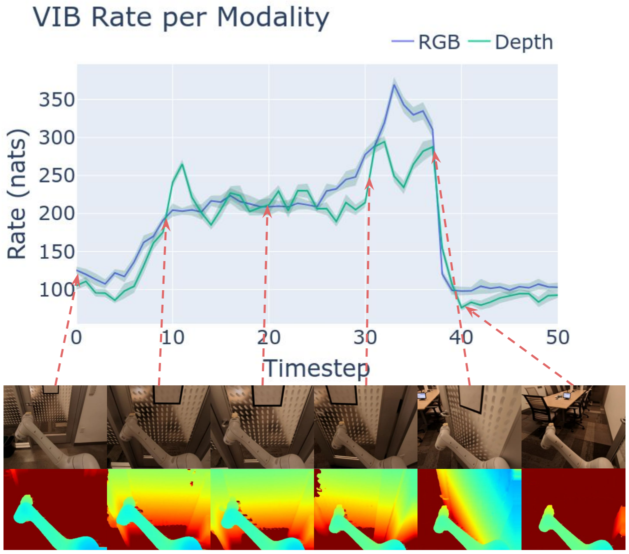


Figure 1. In this work we apply the Variational Information Bottleneck (VIB) to reduce the multi-sensor sim-to-real gap and carry out sensor fusion for a challenging door opening task. The top plot quantifies probabilistic uncertainty of RGB (blue) and Depth (green) sensor modalities as a function of timestep in a sample trajectory, as measured by the VIB rate (divergence of the state representation posterior from its learned marginal). The rate is computed from 8 samples, where the solid line is the mean value and shaded region the standard deviation. The top row of images is RGB observations at each labeled point in time, and the bottom row is the corresponding depth images. In general, we find that the rate is highest when the gripper contacts the door handle or is unlatching the door, for both modalities, suggesting that more information is required for this task-critical phase.

1. Introduction

A long-standing robotics research problem is to develop agents capable of complex behaviours in the real world. One promising approach is to learn from demonstrations (Schaal et al., 1997), where the agent learns by modeling the distribution over expert actions. Within the imitation learning paradigm, behavior cloning (BC) (Bain & Sammut, 1995; Torabi et al., 2018) is a simple supervised learning method for imitating expert behaviors. In spite of well-known shortcomings, such as compounding errors and the inability to surpass expert performance, recent progress in real-world robotics has shown the promise of BC across different domains and tasks (Jang et al., 2022; Florence et al., 2022).

That said, much progress remains to be made towards deploying robots in the real world, especially as we scale to

[1] UC Berkeley, work done while author was at Everyday Robots [2] Everyday Robots [3] Google Research [4] Work done while author was at Google. Correspondence to: Yuqing Du <yuqing_du@berkeley.edu>.

Proceedings of the 39^th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

more unstructured, visually diverse environments such as homes or offices. Learning directly in the real world is often costly and challenging. As a result, many end-to-end learning approaches benefit from training and evaluation in simulation. For example, in this work we use simulated evaluations to determine which BC models are suitable for real-world testing, as each real-world evaluation takes significant time, involves numerous scenarios, and requires hands-on human involvement. However, relying on simulation introduces the well known "reality gap" (Jakobi et al., 1995), where policy performance in simulation does not necessarily transfer to the real world. Many approaches have been proposed for tackling this problem (Tobin et al., 2017; Sadeghi & Levine, 2016), but they can require significant domain knowledge and engineering. This sim-to-real problem is only exacerbated when multiple sensor modalities are used, as they may each have their own "reality gap". Beyond sim-to-real challenges, sensor fusion itself remains an active area of research. Prior work has shown that naive combinations of sensor inputs can hinder policy performance (Huang et al., 2020). Furthermore, the fusion method itself may suffer from a sim-to-real gap, where a technique that successfully fuses simulated modalities may not extrapolate to reality.

Let us consider the desirable properties of a multi-sensor, end-to-end imitation learning policy. The policy should: 1) be invariant to sim and real domains on a per-sensor level, 2) quantify when each sensor representation is ”uncertain” and rely on it less, and 3) be generally applied to combinations of different sensor modalities. We propose that introducing a information bottleneck, specifically, the Variational Information Bottleneck (Alemi et al., 2016), meets these requirements: 1) a bottleneck with finite channel capacity may force sim and real domains to be encoded using shared bits, 2) VIB has been shown to yield calibrated uncertainty estimates to out-of-distribution examples (Alemi et al., 2018), and 3) VIB can be dropped in as an additional layer and loss without modifying the rest of the architecture. To the best of our knowledge, VIB has not been demonstrated to yield calibrated OoD detection capabilities across multiple modalities on real-world robotic systems to date.

We choose a challenging mobile manipulation task in the real world—latched door opening in an unstructured office environment—as our testbed. Door opening is a required capability for any general-purpose mobile robot performing tasks in human environments like homes and offices. Our policy should be able to bridge the sim-to-real gap broadened by constantly changing real office spaces, successfully make use of all available sensor modalities, and handle the combined complexities of manipulation and navigation.

We investigate our hypotheses about the applicability of VIB to improving generalization in large-scale robotic imitation learning as well as sensor fusion. We train a separate stochastic encoder for each sensory input on both sim and real domains, hypothesizing that the VIB objective encourages bits encoding sim and real features to be shared in a domain-agnostic manner, while still being predictive of demonstrated actions. As an ancillary benefit, this approach also admits tractable probabilistic representations of model uncertainty. We find that the learned VIB rates (the KL divergence of the state-posterior from the state-prior, in nats) are useful for estimating which sensor is more reliable, and thus they can be used in a simple softmax-weighted sensor fusion scheme. Our method uses the actions predicted by the modality whose input has the lowest 'model uncertainty'.

Our contributions are as follows:

- We define a behavior cloning approach that uses VIB for learning domain-agnostic embeddings such that we are able to close the sim-to-real gap.
- We demonstrate that our learned embeddings form a meaningful latent space such that the per-instance VIB rate is informative of task-significant inputs, and that the VIB rate can be used as a measure of modality-specific uncertainty for explainable sensor fusion.
- We tackle the challenging robotics problem of latched door opening in unstructured real-world environments, achieving 96% success.

2. Related Work

Information Bottleneck for Control. The information bottleneck (IB) (Tishby et al., 2000) was originally proposed as a method for finding a compressed representation of the input signal that preserves maximum information about the desired output signal. Alemi et al. (2016) extend the IB approach to deep networks by using a variational lower bound of the IB objective and using the reparametrization trick (Kingma & Welling, 2013). While originally studied in the context of image classification, representation learning using the VIB objective has also been applied to other domains such as unsupervised learning (e.g. β-VAE (Higgins et al., 2016)), meta-learning (Du et al., 2020), and control.

In the domain of control, prior work has explored using the IB principle to learn compressed representations. In reinforcement learning, this includes learning task-critical representations (Pacelli & Majumdar, 2020; Lu et al., 2020), improving the stability of actor-critic methods (Igl et al., 2019), tackling the exploration problem (Goyal et al., 2019), and addressing overfitting in offline RL (Kumar et al., 2021). Closer to our work in the context of imitation learning, Peng et al. (2018) propose the Variational Discriminator Bottleneck (VDB) for regularizing the discriminator in adversarial learning methods. While they apply the VDB for adversarial imitation learning in simulation (Ho & Ermon, 2016), our work instead focuses on behavior cloning for real world robotics and avoids the complexities of adversarial training. Rahmatizadeh et al. (2018) use a VAE-GAN to learn end-to-end manipulation from demonstrations, using the image-based reconstruction loss for regularization. Lynch & Sermanet (2021) propose LangLFP, a multicontext conditional VAE-based imitation learning policy that learns a representation invariant to both visual and text modalities. However, they do not explicitly use the uncertainty estimates for modality fusion. To the best of our knowledge, our work is the first visual end-to-end imitation learning work to demonstrate the efficacy of the VIB for both sim-to-real transfer and multi-sensor fusion.

Beyond IB-based approaches, other methods for learning invariant representations for pixel-based control include using image augmentations (Kostrikov et al., 2020), action-image (Khansari et al., 2020), time-contrastive networks (Sermanet et al., 2018), and bisimulation distances (Zhang et al., 2020).

Sim-to-Real. Sim-to-real transfer allows a model trained in a simulated domain to perform well in the real world. This can be accomplished by reducing the reality gap by making a simulation environment more similar to the real world, or, by learning a model representation that is domain-agnostic or robust across many domains.

In this work, we focus on the principle of domain adaptation—in which input from disparate domains are adapted to be more similar— specifically by inducing a domain-agnostic feature representation of the input. The works DANN and DSN (Ganin et al., 2016; Bousmalis et al., 2016) adversarially teach a network to extract features which do not discriminate between sim and real domains. Feature-level adaptation can be conceptually similar to other self-supervised representation learning work, which also aims to increase similarity between embeddings of positive image pairs. These positive pairs have been generated from image augmentations, patches, and color (Chen et al., 2020; Hénaff et al., 2020; Chen & He, 2020; Pathak et al., 2016; Mundhenk et al., 2018; Noroozi & Favaro, 2016; Zhang et al., 2017)—concepts which perturb the input but not ground truth labels, or leverage other invariants based on the input state. (Khansari et al., 2022) proposes a Task Consistency Loss (TCL), which imposes a similarity loss on the embeddings of RetinaGAN (Ho et al., 2021) translated sim-to-real and real-to-sim pairs. Our work is complementary to this method, and we find that sim-to-real transfer is highest when imposing both VIB regularization and the TCL. In this work, we instead rely on an information bottleneck to learn a shared, compressed representation for both simulated and real domains, without requiring image pairs.

Multi-Sensor Fusion. Many prior works have shown that increasing sensor modalities (e.g. tactile, audio, visual) can improve control for domains such as manipulation (Lee et al., 2020) and autonomous navigation (Fayyad et al., 2020). However, how to best combine multiple modalities is an active area of research.

Successful fusion at the representation level for end-to-end imitation learning remains challenging due to causal confusion (Codevilla et al., 2019). The network can learn to ignore or overly focus on a modality due to spurious correlations that can occur when demonstrations are the only supervision signal (Huang et al., 2020). Causal confusion is not limited to the multi-sensor regime; simply increasing the dimensionality of the inputs can also increase the number of spurious correlations. One way to overcome this challenge is to use auxiliary losses to enforce a more meaningful multimodal representation (Xu et al., 2017; Lee et al., 2020); however, developing such auxiliary losses can require task and sensor specific loss engineering.

More similar to our work is a rich literature of approaches that aim to account for sensor uncertainty during fusion. These include probabilistic methods (Proença & Gao, 2018; Murphy, 1998), modeling sensor noise distributions (Zhu et al., 2013), and learned confidence maps (Van Gansbeke et al., 2019). However, these prior approaches are not applied to end-to-end learning. Furthermore, we introduce the insight of using an inherent measure of uncertainty from the VIB for explainable sensor fusion.

3. Preliminaries

Given an input source X, stochastic encoding Z, and target variable Y, the Information Bottleneck (IB) (Tishby et al., 2000) approach optimizes for an encoding Z that is maximally predictive of Y while being a compressed representation of X. A parametric encoder $p(z|x;\theta)$, typically modeled via a neural network with weights θ, is optimized through

$$\max_{\theta} \quad I(Z;Y|\theta) - \beta I(Z;X|\theta) \tag{1}$$

where $I(A;B)$ is the mutual information between variables A, B, formally defined as the KL-divergence between the joint density and the product of the marginals,

$$I(A;B) = D_{KL}[p(a,b)||p(a)p(b)] \tag{2}$$

The hyperparameter β controls the tradeoff between the predictive power and degree of compression of z, where $\beta = 0$ corresponds to a stochastic version of the typical maximum likelihood objective. However, as mutual information is generally computationally intractable, Alemi et al. (2016) propose the Variational Information Bottleneck (VIB) to learn a variational lower bound of the IB, extending the IB method to deep neural networks and high dimensional inputs. The VIB objective to be maximized takes the form:

$$\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{z\sim p(z|x_n)}\left[\log q(y_n|z)-\beta\log\frac{p(z|x_n)}{r(z)}\right] \quad (3)$$

where $p(z|x)$ is our encoder, $q(y|z)$ is a variational approximation to $p(y|z) = \int dx\, p(y|x)p(z|x)p(x)/p(z)$, $r(z)$ is a variational approximation of $p(z) = \int dx\, p(z|x)p(x)$, and N is the number of training examples. We follow the design choice in Alemi et al. (2016) with learning a $p(z|x)$ that is independent of y, as our goal is to use the encoded representation z for a downstream task where y will be unobserved.

As is done in standard practice, we parametrize the encoder distribution $p(z|x)$ as a multivariate Gaussian, where the mean and diagonal terms of the covariance matrix are predicted by a neural network. We use the reparametrization trick to compute derivatives of the network parameters with respect to losses on the stochastic samples. As computing the marginal distribution $p(z)$ exactly is intractable, $r(z)$ is often parametrized by a learned model as well. In our case, we model $r(z)$ as a mixture-of-Gaussians with the number of modes as a hyperparameter.

4. Bayesian Imitation Learning

In imitation learning, we are provided with a set of expert demonstrations, consisting of input observations s and corresponding expert actions a. Our goal is to learn a policy $\pi(a|s)$ that replicates the expert. To make use of the VIB, we decompose the policy into a stochastic encoder, $p(z|s)$, and an action decoder, $q(a|z)$, and we impose a bottleneck on the learned stochastic encoding z using Eq. 3. For a single input image s, we can decompose the training loss as

$$\mathcal{L} = \underbrace{\mathbb{E}_{z\sim p(z|s)}\left[-\log q(a|z)\right]}_{\mathcal{L}_{BC}} + \beta\,\underbrace{\mathbb{E}_{z\sim p(z|s)}\left[\log\frac{p(z|s)}{r(z)}\right]}_{\mathcal{L}_{KL}} \quad (4)$$

where the first term $\mathcal{L}_{BC}$ is the behaviour cloning loss and the second term is the rate $\mathcal{L}_{KL}$, equivalent to $D_{KL}[p(z|x)||r(z)] \geq I(X;Z)$. Following Jang et al. (2022), we use a Huber loss (Huber, 1964) between the predicted and demonstrated actions instead of a negative log-likelihood loss on an explicit distribution. The rate is weighted by β and controls the bottlenecking tradeoff. We use simulated evaluations for the β hyperparameter sweep, as real evaluations are costly and time-consuming. As observed in Lu et al. (2020), in practice we find it helpful to linearly anneal β from 0 within the first 3000 steps of training.

In training our policy with Eq. 4, we aim to learn an encoding space z that is maximally predictive of the expert actions while being maximally concise about the input observations. Since the encoder is trained on both simulated and real images, we hypothesize that the shared yet compressed representations will help close the sim-to-real gap.

Network Details. We parameterize the stochastic encoder $p(z|s)$ using a ResNet-18 (He et al., 2015) that predicts the mean and covariance of a multivariate Gaussian distribution on $\mathbb{R}^{64}$. The action decoder $q(a|z)$ is a 2-layer MLP, and the learned prior $r(z)$ is a mixture of multivariate Gaussians on $\mathbb{R}^{64}$ with 512 components and a learnable mean and diagonal covariance matrix.

We estimate both expectations in Eq. 4 using Monte Carlo sampling. This is necessary for the second term as we parametrize our prior with a multivariate Gaussian mixture model, from which the KL divergence is analytically intractable (Hershey & Olsen, 2007). During training we take 8 samples from the stochastic embedding per input and compute the average VIB rate loss. Similarly, we decode each sample into separate actions and compute the average behavior cloning loss. At inference time, we execute the mean action prediction for a sensor modality by computing the model average across the samples.

4.1. Uncertainty-based Sensor Fusion

Sensor fusion is a well-known challenge in robotics. While increasing the number of sensors increases the amount of information a system can glean from the environment, composing multiple sensor inputs can be problematic, especially if they are in disagreement. Ideally, we would like our multi-sensor system to rely less on inputs it is uncertain about (e.g. dissimilar to the expert data). For example, if the RGB camera records significantly different colors, lights, or objects than those in the training dataset, the model should rely on action predictions from the depth modality instead.

As an estimate of sensor uncertainty, we make use of the per-instance rate $\mathcal{L}_{KL}$. Similarly to Alemi et al. (2018), we expect higher rates to correlate with higher uncertainty about the input, which can also correspond to inputs uncommon in the training dataset (eg. out-of-distribution inputs). Hence we favor the actions from the modalities with lower rates (i.e. more 'in-distribution'), allowing us to blend or switch between sensors at each timestep, in an uncertainty-aware manner.

Given N sensor inputs, we train $i = 1, ..., N$ models independently. That is, each modality's model has its own encoder $p_i(z|x)$, decoder $q_i(a|z)$, and learned marginal $r_i(z)$. This allows us to impose modality-specific bottlenecks, as the ideal lower-dimensional density model may vary across modalities. Each model only sees the input image corresponding to its designated modality. At inference time, we use the VIB rate, $\mathcal{L}^i_{KL}$, of the ith model

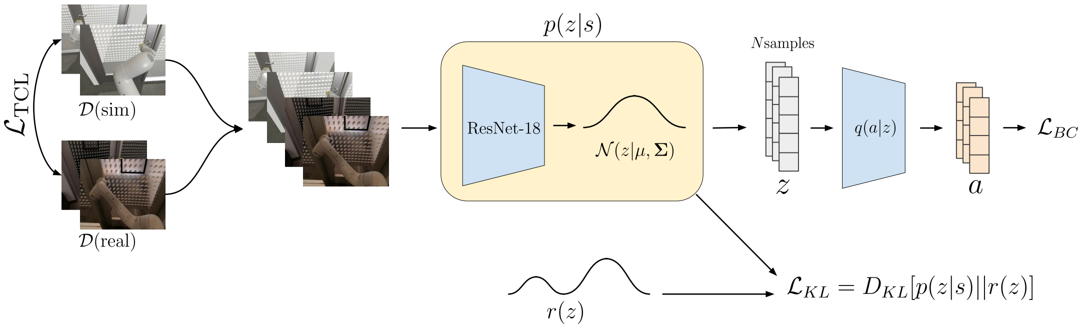


Figure 2. Overall model diagram for a single modality. We learn a stochastic encoder $p(z|s)$, an action decoder $q(a|z)$, and a prior $r(z)$. The objective (Eq. 4) consists of both a behaviour cloning loss and a KL-divergence term for imposing the Variational Information Bottleneck (VIB). We train all components using a mixed sim and real dataset to encourage the latent representation to be domain agnostic. We also mix in GAN-adapted sim-to-real and real-to-sim images among $\mathcal{D}$(sim) and $\mathcal{D}$(real) and impose the proposed Task Consistency Loss (TCL) from Khansari et al. (2022).

as a measure of modality uncertainty. Since we want lower rates to correspond to higher action weights, we first compute the unnormalized weights for the jth model as $\bar{w}^j = \sum_i \mathcal{L}^i_{KL} - \mathcal{L}^j_{KL}$. We then normalize the weights per modality w^i such that $\sum_i w^i = 1$. We experiment with two normalization schemes: 1) softmax normalization for more discrete modality switching, and 2) dividing $\bar{w}^i$ by $\sum_i \mathcal{L}^i_{KL}$ for more blended actions. We compute the fused action as $a = \sum_i w^i a^i$. While we evaluate the two most common image modalities for robotics, RGB and depth, in principle our proposed method can be extended to additional modalities. Figure 3 gives an overview of the fusion process.

5. Experiments

Our experiments aim to answer the following questions: 1) Does the regularization imposed by the information bottleneck lead to domain-agnostic representations, thus closing the sim-to-real gap? 2) Can the VIB-based representation help with making the model more explainable? 3) Can the VIB rates across modalities be used for uncertainty-based sensor fusion?

5.1. Experimental Setup

Our training dataset consists of a real-world dataset of 2068 demonstrations (∼13.5 hours) and a simulated dataset of ∼500 demonstrations (∼2.7 hours), all collected using handheld teleoperation devices. The real-world dataset does not control for the interior of the room, leaving the observations as natural as possible. We also train a RetinaGAN model (Ho et al., 2021) on the sim and real dataset, and use the GAN model to translate sim images to look like real and vice-versa. We use all four datasets (sim, adapted sim, real, and adapted real) to impose a feature-level consistency loss (Khansari et al., 2022) across all models (including baselines).

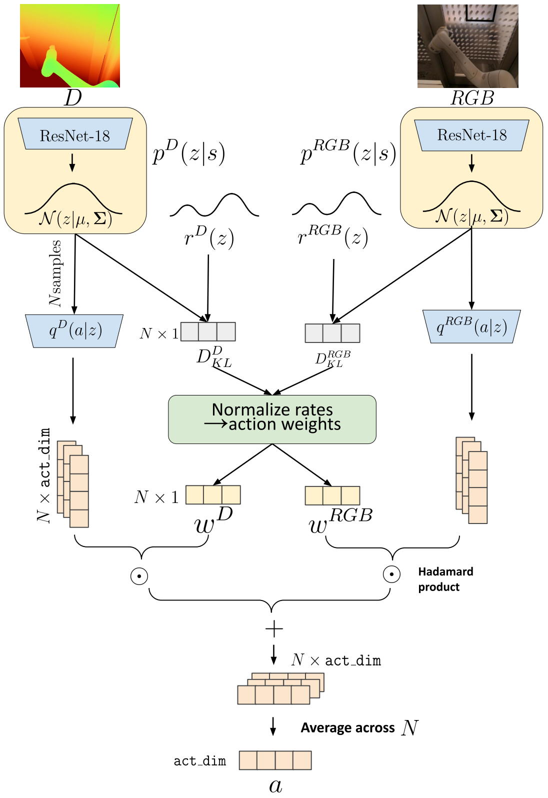


Figure 3. System diagram showing how each modality contributes to the final action prediction. Each sensor is trained to have an independent encoder $p(z|s)$, decoder $q(a|z)$, and prior $r(z)$. During inference, we weigh the contribution of each modality's action prediction using the VIB rate such that modalities with lower rates (correlated with lower uncertainty) are in more control.

We carry out evaluations on 10 different latched doors, split into 5 left swinging and 5 right swinging. Six of the doors are in the training dataset and four are only used during evaluation—entirely unseen during training. Each door is tested with 30 trials split across two different robots, of which only one was used to collect the training dataset. As in the training dataset, we do not control for the interior of the rooms (eg. objects in the room, pose of furniture).

For each model configuration, we train three separate models that differ only by the random training seed. From these, we choose three checkpoints with the highest evaluation performance in simulation to use for real evaluation. During real evaluation, we randomly sample from the three checkpoints in a manner blind to the robot operator. We report estimated standard deviation as $\sqrt{p(1-p)/(n-1)}$, assuming the n trials are i.i.d. Bernoulli variables with success p. For a direct comparison, we follow the experimental procedure in (Khansari et al., 2022).

5.2. Results and Discussion

Q1. Domain-Agnostic Representations. Since we use the same encoder for both simulated and real images of the same modality, our goal is to learn a bottlenecked encoding that is domain agnostic, i.e. helps close the sim-to-real gap. Since our approach enables us to learn a lower-dimensional density model of the input images, we can directly investigate whether this is the case. To do so, we measure the KL divergence between the distributions parametrized by the embeddings of simulated and real images.

In Figure 4, we look at the nearest neighbours, measured by KL divergence, within a subset of our training dataset consisting of 1600 mixed sim and real images. We look at three primary phases of the task: 1) approaching the door, 2) unlatching and opening the door, and 3) entering the room. Both (1) and (3) only focus on base motion for navigation, while (2) requires a mixture of base and arm motion for manipulation. As (3) includes the interior of the rooms, images from this phase also have greater visual diversity.

In general, we find that the closest images in the embedding space within a KL divergence of $\sim$400 nats correspond to similar actions, which generally corresponds to phases of the door opening task. Notably, the nearest neighbours of simulated images include real images and vice versa, suggesting that our learned representations do not separate images from the two domains when they correspond to similar actions—thus closing the sim-to-real gap. For the corresponding sim-to-real gap investigations in the depth modality, see Appendix B.

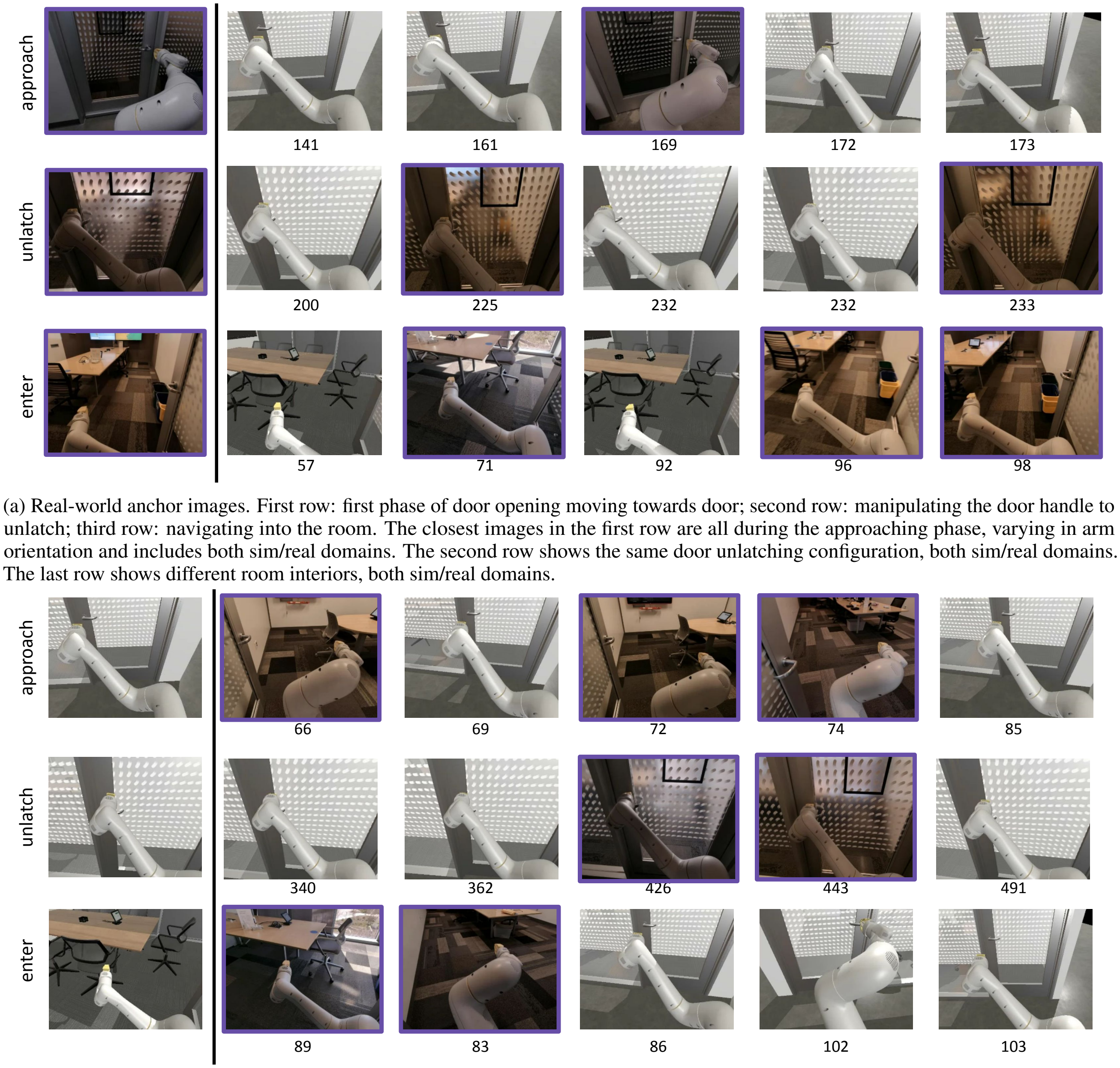
The first and last rows of Figure 4b show an example where the closest images may correspond to a different phase than the anchor image—images anchored by approaching the door show nearest neighbours that are entering the room and vice versa. However, when inspecting the network action predictions for these instances, we find that the actions correspond to the same commands of moving the base forward and keeping the arm still. This aligns well with our expectation from the model: the IB objective aims to discard everything that is not relevant for the auxiliary variable—action prediction in our case. The nearest neighbours showing visually different images but corresponding to the same actions suggests that the embedding is discarding anything that is not predictive of the actions.

Q2. Model Explainability. During training, Eq. 4 aims to minimize the average rate $\mathcal{L}_{KL}$, or KL divergence between the embedding distribution and the learned marginal, to a budget determined by the hyperparameter β. In doing so, the model must allocate how much divergence is acceptable for each datapoint, which can vary nonuniformly across the dataset. Inputs that are common or correspond to easily predictable actions require less information to encode (lower rates), while the opposite is true for uncommon or hard-to-compress inputs (higher rates).

We hypothesize that the VIB rate should be higher during portions of the task that are more challenging, as these states likely require more information to act accurately. To investigate this, we plot the change in rate across trajectories for both modalities. A sample is shown in Figure 1, with additional trajectories of different conditions in Appendix C. In general, we find that the rates for both modalities are highest during the most challenging part of the task—manipulating the door latch. This phase is critical, as to successfully open the door, the agent should: control both the arm and the base (9 degree-of-freedom), know which side of the handle to push down on (depending on swing orientation), and take care not to slip off the handle. On the other hand, the rate is lowest for the most intuitively compressible inputs—where the robot only needs to move the base at the start and end of the task (2 degrees-of-freedom). This suggests that the learned prior and encoder are salient to the challenging parts of the task, making the model more explainable.

Furthermore, as shown in Figure 4, finding the nearest neighbours between a set of anchor images within a subset of the dataset can be used to provide more insights about the quality of a trained model. For example, wrong groupings of images could indicate the bottleneck was too tight for the model to properly distinguish the different phases of the task; or, retrieving both sim and real images for a given anchor image is a healthy sign that the model has learned a domain-agnostic representation.

Q3. VIB Rate-based Sensor Fusion. The previous two experiments suggest that we are capable of 1) learning a representation that helps close the sim-to-real gap, and 2) using the VIB rates to better understand and debug the trained model post hoc. Here, we test our hypotheses that both the reduced sim-to-real gap and VIB rate-based sensor fusion help performance on the real-world task of latched



(a) Real-world anchor images. First row: first phase of door opening moving towards door; second row: manipulating the door handle to unlatch; third row: navigating into the room. The closest images in the first row are all during the approaching phase, varying in arm orientation and includes both sim/real domains. The second row shows the same door unlatching configuration, both sim/real domains. The last row shows different room interiors, both sim/real domains.

(b) Simulation anchor images. First row: first phase of door opening moving towards door; second row: manipulating the door handle to unlatch; third row: navigating into the room. Interestingly, the closest images in the first and last rows both show sim/real images in either the approaching phase or navigating phase within the room, likely due to the similar base movement and lack of arm motion. The second row shows similar door unlatching frames across sim/real domains.

Figure 4. Anchor images (real) displayed in the leftmost column, with the nearest 5 images to their right. KL divergences (measured relative to the anchor image) given below each similar image. Purple borders indicate real images.

door opening. As observed in other sensor fusion works and our own baselines, combining modalities can sometimes be more detrimental than using either modality individually.

In Table 1, we compare our method against three baseline models from Khansari et al. (2022): two trained on each modality individually and one fusion model that concatenates RGB and depth embeddings together before feeding into the action prediction MLP, a common multi-sensor fusion approach (Park et al., 2017; Calandra et al., 2018). Each baseline has the same architecture as ours, minus the stochastic embedding. All baselines and our methods make use of the CycleGAN-based adaptation and sim-real Task Consistency Loss (TCL) proposed in Khansari et al. (2022).

We evaluate three fusion methods with VIB: 1) embedding

Concatenation Fusion (CF), 2) VIB rate-based action fusion with linear normalization (blended actions), and 3) VIB rate-based action switching with softmax normalization (modality switching). We find that the domain agnostic representations induced by the bottleneck help improve performance for all three VIB methods above the baselines. In particular, the best VIB methods outperform the best baseline TCL-only method (RGB, Table 1 row 1) by 16%, and the best baseline TCL-only fusion method by 21%. Notably, baselines include TCL and RetinaGAN, showing that VIB is composable modularly with other domain adaptation methods for improved transfer. In the last row of Table 1, we also report the performance of the best VIB fusion method alone without TCL, where we find comparable performance to the baseline TCL methods. This suggests that the TCL and VIB individually improve transfer by similar amounts, and we receive the most gains from composing the two methods.

Interestingly, we find that the RGBD + VIB (CF) method performs equally well as VIB (Softmax Fusion), even though RGBD (CF) was the lowest performing baseline in the real evaluations. However, we also find that the CF method has much higher variance in simulated evaluation performance, suggesting that relying on concatenated embeddings can increase optimization difficulty. See Appendix D for corresponding plots. We note that the softmax-based fusion method performs just as well, while additionally having a more interpretable sensor fusion policy by construction. Examining the rates per modality and using the rates to enforce which modality is in control gives assurances as to which sensor the policy is relying on at each point in time, allowing for easier model debugging. Notably, the black box CF policy cannot provide such insights.

Between the rate-based fusion methods, we find that softmax normalization performs slightly better than linear normalization. We hypothesize that this is due to the discrete action selection, whereas the linear method blends the actions using linear interpolation. This may be problematic if each modality has learned to predict different actions for the same state, and the blended action is worse than either action individually. For this particular task, we find that switching modalities works better than blending them.

To investigate whether the rate-based sensor fusion is correctly choosing the ‘better’ modality (i.e. that the better performing modality individually has lower rates), we ablate the softmax-based fusion model into its RGB and depth only components in Table 2. Using the same checkpoints, we enforce that the action taken is always from one of the modalities. Interestingly, the depth-only branch performs slightly better than the fused model, while the RGB-only branch performs much worse. Our fused model’s performance is close to the depth-only result, suggesting that the rate-based fusion is able to pick out the stronger performing modality and avoid relying on the lower performing modality. We note that the RGB performance here is lower than the baselines due to the checkpoints being originally selected for best softmax-fusion performance in simulation, not best RGB-only performance.

Method	Total	Seen	Unseen
RGB	80% ± 2.3	75% ± 3.2	87% ± 3.1
Depth	77% ± 2.5	79% ± 3.1	75% ± 4.2
RGBD (CF)	75% ± 2.4	79% ± 3.0	69% ± 4.3
RGBD + VIB (CF)	96% ± 1.1	94% ± 1.8	98% ± 1.3
RGBD + VIB (Linear)	93% ± 1.5	93% ± 1.9	93% ± 2.3
RGBD + VIB (Softmax)	96% ± 1.1	98% ± 1.0	94% ± 2.2
VIB (Softmax) - TCL	75% ± 2.4	82% ± 2.9	65% ± 4.4

Table 1. Real-world door opening success rates (%) ± standard deviation, based on 300 trials (180 on seen doors, 120 on unseen doors). We compare against baseline models that use each modality individually and a concatenation fusion (CF) model that fuses RGB and depth by concatenating the embeddings together before passing to a shared MLP action decoder. For the VIB models, we compare a CF variant that concatenates bottlenecked embeddings and two rate-based fusion variants using the action fusion schemes described in Section 4.1. We find that both softmax fusion and CF perform best, with the former having the additional benefit of an explicitly understandable fusion scheme. In the last row we also report an ablation of the best rate-based fusion method by removing the TCL. The success rates are comparable to the results we get with concatenated RGBD + TCL only (row 3), suggesting that VIB or TCL *alone* have comparable performance in the multimodal setting.

Method	Total	Seen	Unseen
RGBD + VIB Softmax	96% ± 1.1	98% ± 1.0	94% ± 2.2
RGB + VIB	74% ± 2.5	79% ± 3.0	67% ± 4.3
Depth + VIB	98% ± .06	99% ± .07	97% ± 1.6

Table 2. Sensor modality ablation for the softmax-normalized VIB rate-based fusion model. We decompose the contributions of each modality to the Fusion model by using the same model checkpoints, but forcing the model to place all weight on either RGB or Depth. We find that the fused model’s performance is closer to the better performing depth-only model, suggesting that the rate-based fusion is able to select the stronger performing modality.

6. Conclusions and Future Work

Motivated by the challenges of closing the multi-sensor sim-to-real gap in end-to-end learning for robotics, we make use of the regularizing capabilities of the VIB to learn a domain agnostic representation. Compared to other common regularization approaches for sim-to-real (eg. domain randomization), the VIB requires no simulation engineering, instead requiring tuning the hyperparameter β to find a desirable bottleneck capacity. To combine multiple sensor modalities, our insight is to leverage the VIB’s inherent uncertainty estimation for uncertainty-based sensor fusion. Studying the mobile manipulation task of latched door opening in a real office, we highlight some explainability characteristics of our approach, finding that the rates are salient and that the learned embeddings are sim and real domain agnostic. We evaluate and discuss the tradeoffs between different methods of sensor fusion, significantly improving task performance and successfully deploying the VIB on a real multi-sensor robotic system.

Limitations. In this paper we assume policy uncertainty correlates with the VIB rates, since the former is a harder metric to measure. Under the VIB objective, the encoder is given an average rate budget to allocate across all of the inputs seen. Generally, we should expect common inputs will be economically represented by low rates. High rate inputs are inputs that the encoder has difficulty compressing. While it seems reasonable to, and our results indicate that it is appropriate to treat these as examples the policy will perform worse on, this does rely on the encoder and marginal pair being well-aligned. One possible way to mitigate this issue is to use the tools illustrated in Figures 1 and 4 to examine the quality of the model more directly.

Broader Impact. By learning an explicit (albeit high-dimensional) density model of data, we have the tools to better understand what a policy has learned (e.g. whether domain-agnostic images from similar portions of the task are similarly encoded under the stochastic embedding). Using the VIB rate as a measure of sensor uncertainty allows the policy to enforce which modality is in control, leading to a more interpretable understanding of which inputs the policy uses to make its decisions. Our hope is that such a capability improves the safety and reliability of robotic systems trained end-to-end with machine learning. Although door opening is a benign application of robotics in and of itself, a potential outcome of this capability is that it unlocks a myriad of indoor robotics applications for which navigating autonomously between rooms was previously very challenging, such as security and patrol inside buildings.

Future work. In this work, we find that the regularization provided by a bottleneck objective helps to reduce the sim-to-real gap for robotics applications. Other commonly used regularization approaches include using contrastive losses (to learn invariant representations) and/or data augmentations. For future work, it would be interesting to thoroughly investigate and understand the interplay between these types of regularization—whether they provide complementary or redundant regularization capabilities, with what tradeoffs, and how their utilities may vary across learning domains and applications.

References

Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Alemi, A. A., Fischer, I., and Dillon, J. V. Uncertainty in the variational information bottleneck. *arXiv preprint arXiv:1807.00906*, 2018.

Bain, M. and Sammut, C. A framework for behavioural cloning. In *Machine Intelligence 15*, pp. 103–129, 1995.

Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., and Erhan, D. Domain separation networks. *Advances in neural information processing systems*, 29:343–351, 2016.

Calandra, R., Owens, A., Jayaraman, D., Lin, J., Yuan, W., Malik, J., Adelson, E. H., and Levine, S. More than a feeling: Learning to grasp and regrasp using vision and touch. *IEEE Robotics and Automation Letters*, 3(4): 3300–3307, 2018.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Chen, X. and He, K. Exploring simple siamese representation learning, 2020.

Codevilla, F., Santana, E., López, A. M., and Gaidon, A. Exploring the limitations of behavior cloning for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9329–9338, 2019.

Du, Y., Xu, J., Xiong, H., Qiu, Q., Zhen, X., Snoek, C. G., and Shao, L. Learning to learn with variational information bottleneck for domain generalization. In *European Conference on Computer Vision*, pp. 200–216. Springer, 2020.

Fayyad, J., Jaradat, M. A., Gruyer, D., and Najjaran, H. Deep learning sensor fusion for autonomous vehicle perception and localization: A review. *Sensors*, 20(15):4220, 2020.

Florence, P., Lynch, C., Zeng, A., Ramirez, O. A., Wahid, A., Downs, L., Wong, A., Lee, J., Mordatch, I., and Tompson, J. Implicit behavioral cloning. In *Conference on Robot Learning*, pp. 158–168. PMLR, 2022.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Goyal, A., Islam, R., Strouse, D., Ahmed, Z., Botvinick, M., Larochelle, H., Bengio, Y., and Levine, S. Infobot: Transfer and exploration via the information bottleneck. *arXiv preprint arXiv:1901.10902*, 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition, 2015.

Hershey, J. R. and Olsen, P. A. Approximating the kullback leibler divergence between gaussian mixture models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pp. IV–317. IEEE, 2007.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

Ho, D., Rao, K., Xu, Z., Jang, E., Khansari, M., and Bai, Y. Retinagan: An object-aware approach to sim-to-real transfer. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10920–10926, 2021. doi: 10.1109/ICRA48506.2021.9561157.

Ho, J. and Ermon, S. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29:4565–4573, 2016.

Huang, Z., Lv, C., Xing, Y., and Wu, J. Multi-modal sensor fusion-based deep neural network for end-to-end autonomous driving with scene understanding. *IEEE Sensors Journal*, 21(10):11781–11790, 2020.

Huber, P. J. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35:73–101, 1964.

Hénaff, O. J., Srinivas, A., Fauw, J. D., Razavi, A., Doersch, C., Eslami, S. M. A., and van den Oord, A. Data-efficient image recognition with contrastive predictive coding, 2020.

Igl, M., Ciosek, K., Li, Y., Tschiatschek, S., Zhang, C., Devlin, S., and Hofmann, K. Generalization in reinforcement learning with selective noise injection and information bottleneck. *arXiv preprint arXiv:1910.12911*, 2019.

Jakobi, N., Husbands, P., and Harvey, I. Noise and the reality gap: The use of simulation in evolutionary robotics. In *European Conference on Artificial Life*, pp. 704–720. Springer, 1995.

Jang, E., Irpan, A., Khansari, M., Kappler, D., Ebert, F., Lynch, C., Levine, S., and Finn, C. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning*, pp. 991–1002. PMLR, 2022.

Khansari, M., Kappler, D., Luo, J., Bingham, J., and Kalakrishnan, M. Action image representation: Learning scalable deep grasping policies with zero real world data. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3597–3603, 2020.

Khansari, M., Ho, D., Du, Y., Fuentes, A., Bennice, M., Sievers, N., Kirmani, S., Bai, Y., and Jang, E. Practical imitation learning in the real world via task consistency loss, 2022.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kostrikov, I., Yarats, D., and Fergus, R. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. *arXiv preprint arXiv:2004.13649*, 2020.

Kumar, A., Singh, A., Tian, S., Finn, C., and Levine, S. A workflow for offline model-free robotic reinforcement learning. *arXiv preprint arXiv:2109.10813*, 2021.

Lee, M. A., Zhu, Y., Zachares, P., Tan, M., Srinivasan, K., Savarese, S., Fei-Fei, L., Garg, A., and Bohg, J. Making sense of vision and touch: Learning multimodal representations for contact-rich tasks. *IEEE Transactions on Robotics*, 36(3):582–596, 2020.

Lu, X., Lee, K., Abbeel, P., and Tiomkin, S. Dynamics generalization via information bottleneck in deep reinforcement learning. *arXiv preprint arXiv:2008.00614*, 2020.

Lynch, C. and Sermanet, P. Language conditioned imitation learning over unstructured data. *Proceedings of Robotics: Science and Systems. doi*, 10, 2021.

Mundhenk, T. N., Ho, D., and Chen, B. Y. Improvements to context based self-supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9339–9348, 2018.

Murphy, R. R. Dempster-shafer theory for sensor fusion in autonomous mobile robots. *IEEE Transactions on robotics and automation*, 14(2):197–206, 1998.

Noroozi, M. and Favaro, P. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pp. 69–84. Springer, 2016.

Pacelli, V. and Majumdar, A. Learning task-driven control policies via information bottlenecks. *arXiv preprint arXiv:2002.01428*, 2020.

Park, S.-J., Hong, K.-S., and Lee, S. Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pp. 4980–4989, 2017.

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

Peng, X. B., Kanazawa, A., Toyer, S., Abbeel, P., and Levine, S. Variational discriminator bottleneck: Improving imitation learning, inverse rl, and gans by constraining information flow. *arXiv preprint arXiv:1810.00821*, 2018.

Proença, P. F. and Gao, Y. Probabilistic rgb-d odometry based on points, lines and planes under depth uncertainty. *Robotics and Autonomous Systems*, 104:25–39, 2018.

Rahmatizadeh, R., Abolghasemi, P., Bölöni, L., and Levine, S. Vision-based multi-task manipulation for inexpensive robots using end-to-end learning from demonstration. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 3758–3765. IEEE, 2018.

Sadeghi, F. and Levine, S. Cad2rl: Real single-image flight without a single real image. *arXiv preprint arXiv:1611.04201*, 2016.

Schaal, S. et al. Learning from demonstration. *Advances in neural information processing systems*, pp. 1040–1046, 1997.

Sermanet, P., Lynch, C., Chebotar, Y., Hsu, J., Jang, E., Schaal, S., Levine, S., and Brain, G. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 1134–1141. IEEE, 2018.

Tishby, N., Pereira, F. C., and Bialek, W. The information bottleneck method, 2000.

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., and Abbeel, P. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 23–30. IEEE, 2017.

Torabi, F., Warnell, G., and Stone, P. Behavioral cloning from observation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 4950–4957, 2018.

Van Gansbeke, W., Neven, D., De Brabandere, B., and Van Gool, L. Sparse and noisy lidar completion with rgb guidance and uncertainty. In *2019 16th international conference on machine vision applications (MVA)*, pp. 1–6. IEEE, 2019.

Xu, H., Gao, Y., Yu, F., and Darrell, T. End-to-end learning of driving models from large-scale video datasets. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2174–2182, 2017.

Zhang, A., McAllister, R., Calandra, R., Gal, Y., and Levine, S. Learning invariant representations for reinforcement learning without reconstruction. *arXiv preprint arXiv:2006.10742*, 2020.

Zhang, R., Isola, P., and Efros, A. A. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1058–1067, 2017.

Zhu, H., Leung, H., and He, Z. A variational bayesian approach to robust sensor fusion based on student-t distribution. *Information Sciences*, 221:201–214, 2013.

A. Detailed Experiment Results

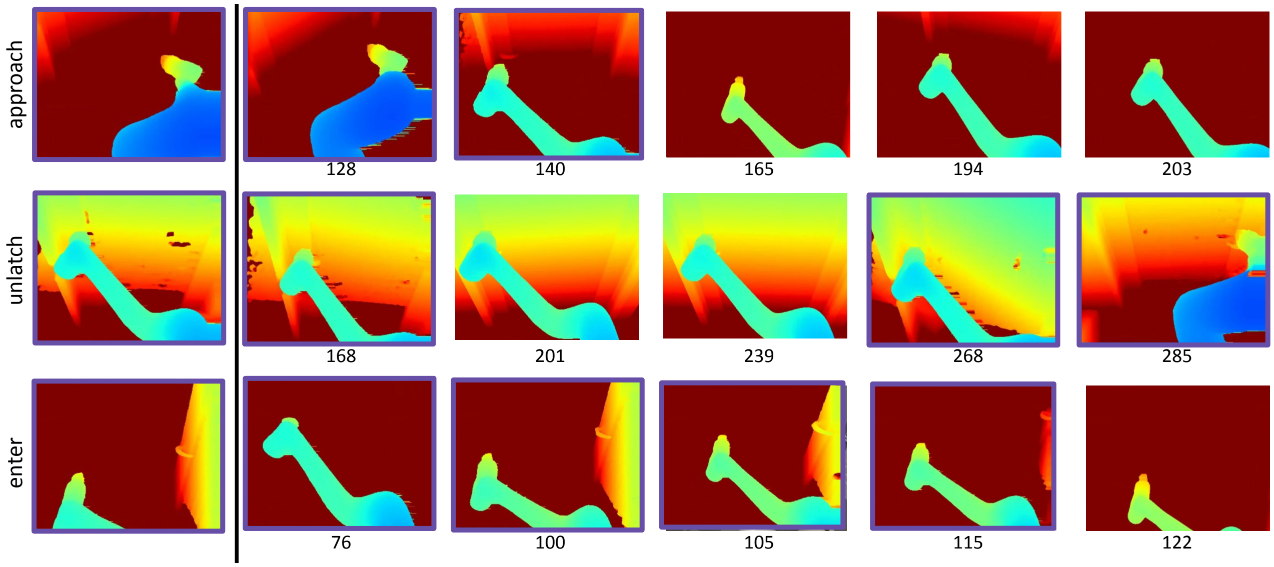
	Overall	TL1		TL2		TR1		TR2		TL3		TR3	
Robot		A	B	A	B	A	B	A	B	A	B	A	B
Time of Day		○	◐	◐	●	○	◐	◐	●	●	○	●	○
Lighting		On	Off	Off	On	On	Off	Off	On	On	Off	Off	On
RGB	75%	100%	73%	47%	73%	93%	7%	87%	100%	73%	80%	80%	87%
Depth	79%	80%	100%	60%	100%	87%	53%	100%	53%	27%	100%	100%	87%
RGBD (CF)	79%	53%	93%	40%	100%	100%	40%	100%	100%	33%	93%	100%	100%
RGBD + VIB (CF)	94%	87%	100%	73%	100%	87%	100%	100%	100%	100%	100%	87%	93%
RGBD + VIB (Linear)	93%	80%	100%	87%	93%	93%	100%	100%	100%	80%	100%	80%	100%
RGBD + VIB (Softmax)	98%	100%	100%	100%	93%	100%	100%	93%	100%	100%	100%	87%	100%

Table 3. Full results for training doors, broken down by robot, time of day [○ Morning, ◐ Noon, ● Afternoon], and lighting conditions.

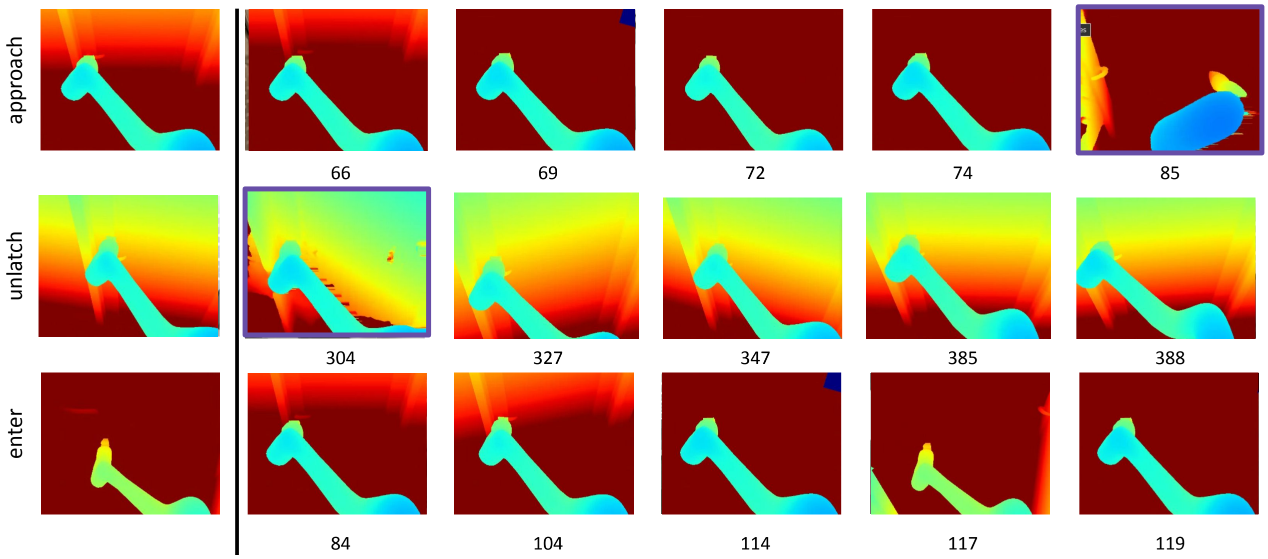
	Overall	ER1		EL1		EL2		ER2	
Robot		A	B	A	B	A	B	A	B
Time of Day		○	●	○	●	○	◐	◐	●
Lighting		On	Off	Off	On	On	Off	Off	On
RGB	87%	100%	100%	47%	80%	80%	93%	100%	93%
Depth	75%	100%	47%	60%	60%	40%	93%	100%	100%
RGBD (CF)	69%	93%	47%	53%	40%	87%	67%	93%	73%
RGBD + VIB (CF)	98%	100%	93%	100%	100%	93%	100%	100%	100%
RGBD + VIB (Linear)	93%	100%	100%	80%	100%	93%	80%	100%	93%
RGBD + VIB (Softmax)	93%	100%	93%	87%	100%	100%	100%	80%	80%

Table 4. Full results for evaluation doors, broken down by robot, time of day [○ Morning, ◐ Noon, ● Afternoon], and lighting conditions.

B. Depth Sim-to-Real Gap



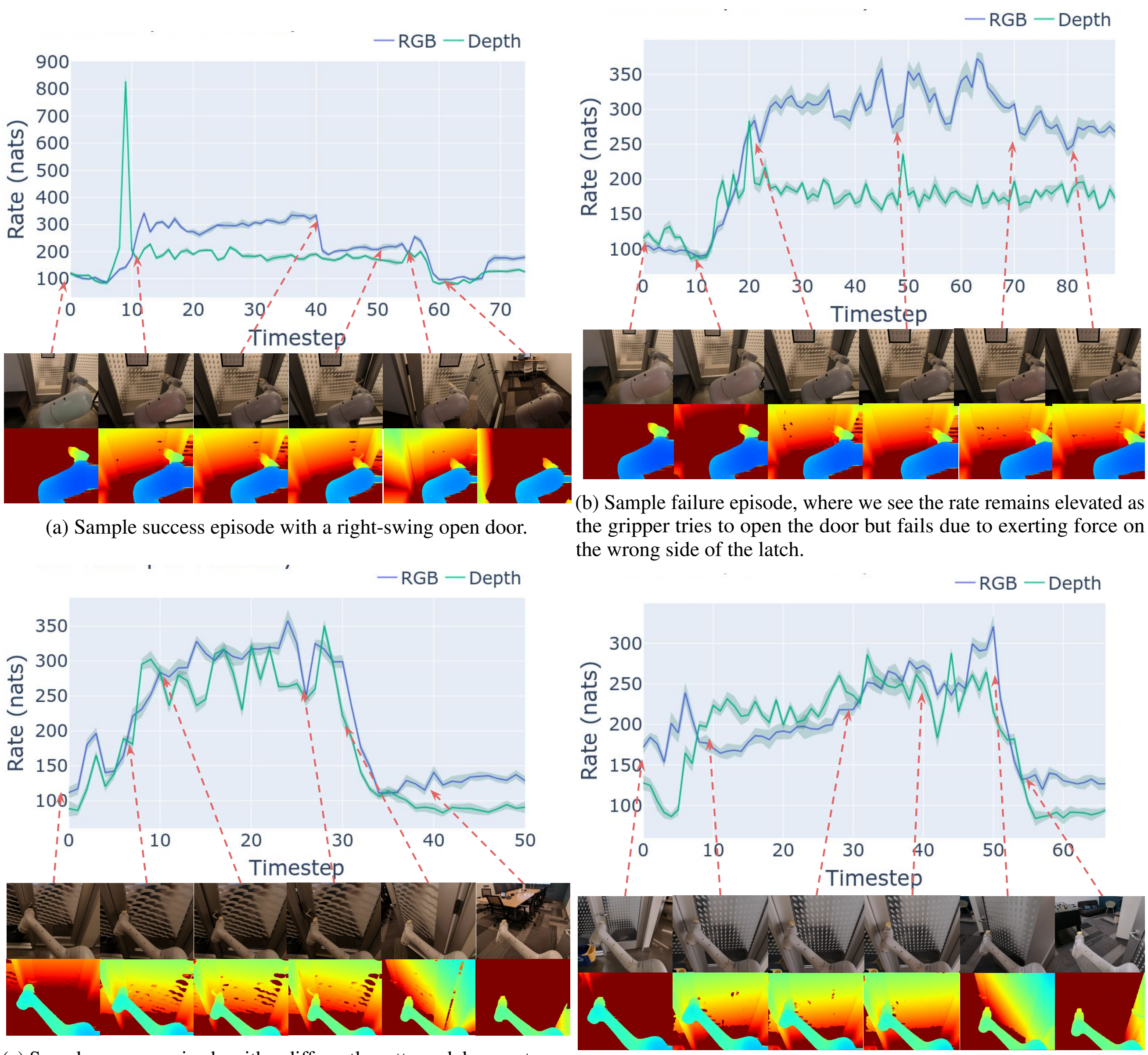
(a) First row: first phase of door opening moving towards door, second row: manipulating door handle to unlatch, third row: navigating into the room. Real anchor images. The closest images in the first row are all during the approaching phase, varying in arm orientation and showing both sim/real domains. The second row shows the same door unlatching configuration, both sim/real domains. The last row shows different room interiors, both sim/real domains. Purple outline indicates real depth images.



(b) First row: first phase of door opening moving towards door, second row: manipulating door handle to unlatch, third row: navigating into the room. Simulated anchor images. Interestingly, the closest images in the first and last rows both show sim/real images in either the approaching phase or navigating phase within the room, likely due to the similar base movement. The second row shows similar door unlatching frames across sim/real domains. Purple outline indicates real depth images.

Figure 5. Anchor image (real) on the left, with the nearest 5 images on the right. KL divergences (measured relative to the anchor image) given below each similar image.

C. Additional Rate Trajectories



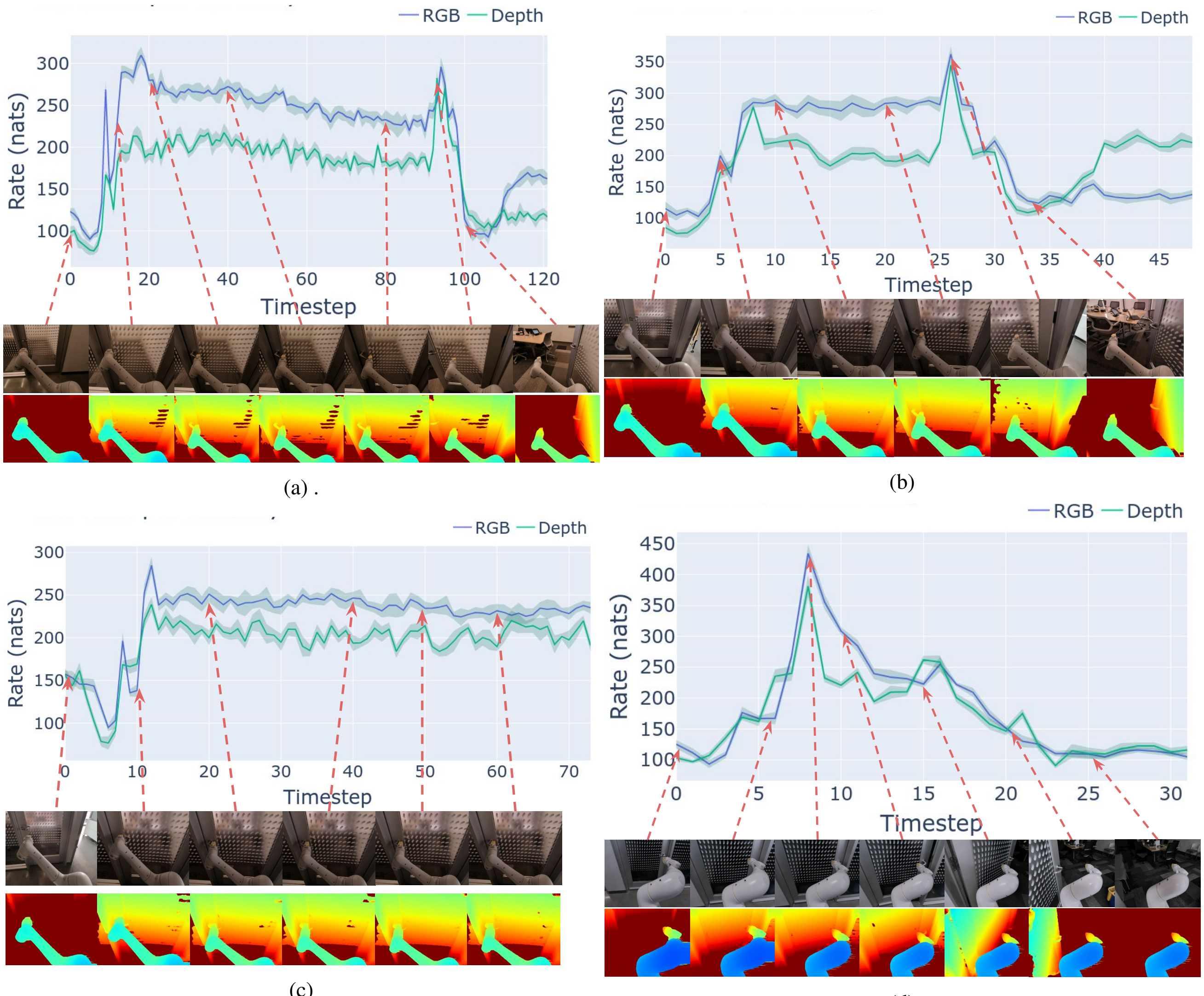
(a) Sample success episode with a right-swing open door.

(b) Sample failure episode, where we see the rate remains elevated as the gripper tries to open the door but fails due to exerting force on the wrong side of the latch.

(c) Sample success episode with a differently patterned door, not seen during training.

(d) Sample success episode with a differently coloured interior.

Figure 6. Rates per modality across additional sample trajectories. Plots show VIB Rate (nats) over time, with corresponding RGB (blue) and depth (green) images labelled with red arrows. In general, the rates are highest during the critical door unlatching phase.



(a) .

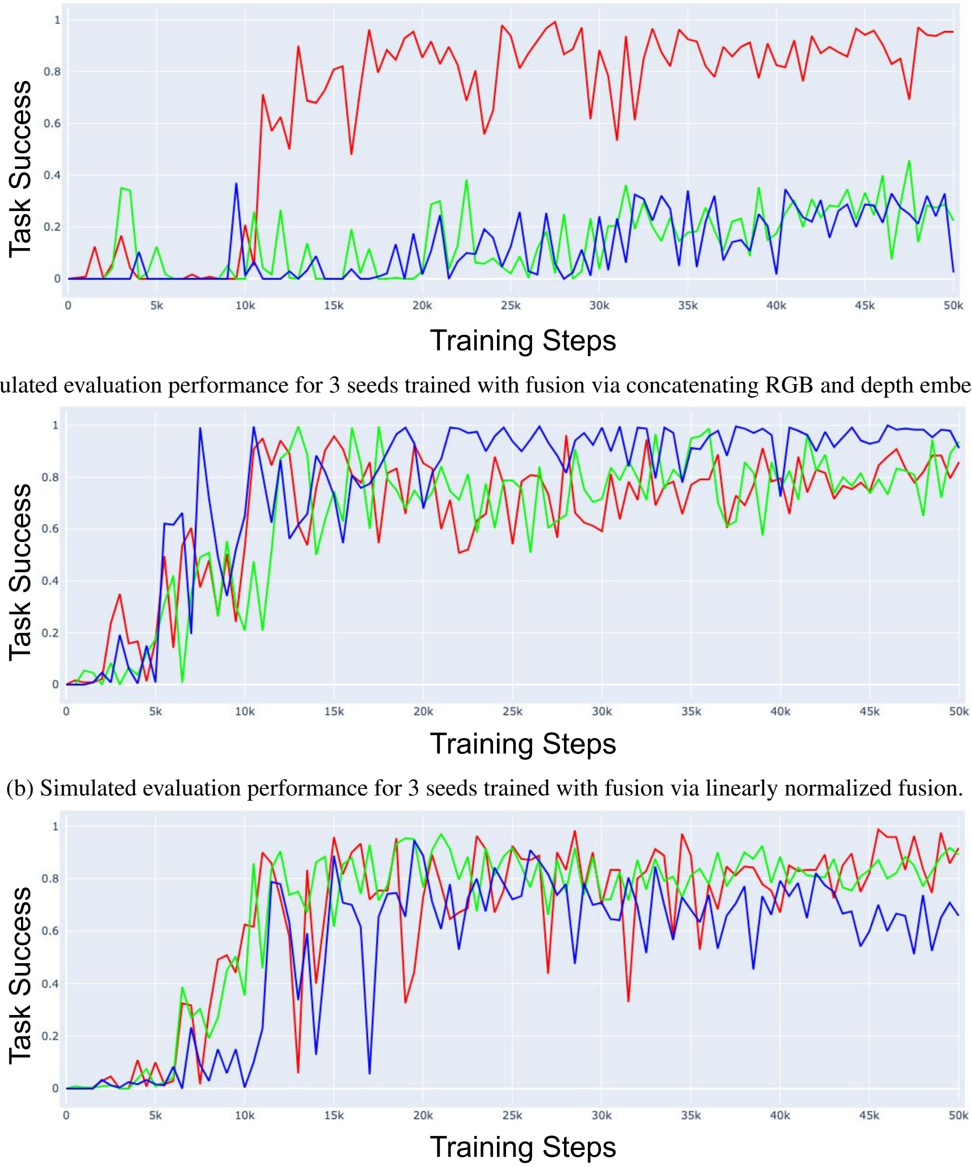
(b)

(c)

(d)

Figure 7. Rates per modality across additional sample trajectories. Plots show VIB Rate (nats) over time, with corresponding RGB (blue) and depth (green) images labelled with red arrows. In general, the rates are highest during the critical door unlatching phase.

D. Simulated Evaluations



(a) Simulated evaluation performance for 3 seeds trained with fusion via concatenating RGB and depth embeddings.

(b) Simulated evaluation performance for 3 seeds trained with fusion via linearly normalized fusion.

(c) Simulated evaluation performance for 3 seeds trained with fusion via softmax normalized fusion.

Figure 8. Simulated evaluation success rates on 6 seen simulated rooms throughout training. For real evaluations we choose the highest performing checkpoints from sim. Notably, the concatenated fusion model is harder to train, with greater variance between seeds.