
Embed to Control: A Locally Linear Latent Dynamics Model for Control from Raw Images

Manuel Watter* **Jost Tobias Springenberg***

Joschka Boedecker

University of Freiburg, Germany

{watterm, springj, jboedeck}@cs.uni-freiburg.de

Martin Riedmiller

Google DeepMind

London, UK

riedmiller@google.com

Abstract

We introduce Embed to Control (E2C), a method for model learning and control of non-linear dynamical systems from raw pixel images. E2C consists of a deep generative model, belonging to the family of variational autoencoders, that learns to generate image trajectories from a latent space in which the dynamics is constrained to be locally linear. Our model is derived directly from an optimal control formulation in latent space, supports long-term prediction of image sequences and exhibits strong performance on a variety of complex control problems.

1 Introduction

Control of non-linear dynamical systems with continuous state and action spaces is one of the key problems in robotics and, in a broader context, in reinforcement learning for autonomous agents. A prominent class of algorithms that aim to solve this problem are model-based locally optimal (stochastic) control algorithms such as iLQG control [1, 2], which approximate the general non-linear control problem via local linearization. When combined with receding horizon control [3], and machine learning methods for learning approximate system models, such algorithms are powerful tools for solving complicated control problems [3, 4, 5]; however, they either rely on a known system model or require the design of relatively low-dimensional state representations. For real *autonomous* agents to succeed, we ultimately need algorithms that are capable of controlling complex dynamical systems from *raw sensory input (e.g. images) only*. In this paper we tackle this difficult problem.

If stochastic optimal control (SOC) methods were applied directly to control from raw image data, they would face two major obstacles. First, sensory data is usually high-dimensional – i.e. images with thousands of pixels – rendering a naive SOC solution computationally infeasible. Second, the image content is typically a highly non-linear function of the system dynamics underlying the observations; thus model identification and control of this dynamics are non-trivial.

While both problems could, in principle, be addressed by designing more advanced SOC algorithms we approach the “optimal control from raw images” problem differently: turning the problem of locally optimal control in high-dimensional non-linear systems into one of identifying a low-dimensional latent state space, in which locally optimal control can be performed robustly and easily. To learn such a latent space we propose a new deep generative model belonging to the class of variational autoencoders [6, 7] that is derived from an iLQG formulation in latent space. The resulting *Embed to Control (E2C)* system is a probabilistic generative model that holds a belief over viable trajectories in sensory space, allows for accurate long-term planning in latent space, and is trained fully unsupervised. We demonstrate the success of our approach on four challenging tasks for control from raw images and compare it to a range of methods for unsupervised representation learning. As an aside, we also validate that deep up-convolutional networks [8, 9] are powerful generative models for large images.

*Authors contributed equally.

2 The Embed to Control (E2C) model

We briefly review the problem of SOC for dynamical systems, introduce approximate locally optimal control in latent space, and finish with the derivation of our model.

2.1 Problem Formulation

We consider the control of unknown dynamical systems of the form

$$\mathbf{s}_{t+1} = f(\mathbf{s}_t, \mathbf{u}_t) + \boldsymbol{\xi}, \quad \boldsymbol{\xi} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\xi}), \quad (1)$$

where t denotes the time steps, $\mathbf{s}_t \in \mathbb{R}^{n_s}$ the system state, $\mathbf{u}_t \in \mathbb{R}^{n_u}$ the applied control and $\boldsymbol{\xi}$ the system noise. The function $f(\mathbf{s}_t, \mathbf{u}_t)$ is an arbitrary, smooth, system dynamics. We equivalently refer to Equation (1) using the notation $P(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{u}_t)$, which we assume to be a multivariate normal distribution $\mathcal{N}(f(\mathbf{s}_t, \mathbf{u}_t), \boldsymbol{\Sigma}_{\xi})$. We further assume that we are only given access to visual depictions $\mathbf{x}_t \in \mathbb{R}^{n_x}$ of state \mathbf{s}_t . This restriction requires solving a joint state identification and control problem. For simplicity we will in the following assume that \mathbf{x}_t is a fully observed depiction of \mathbf{s}_t , but relax this assumption later.

Our goal then is to infer a low-dimensional latent state space model in which optimal control can be performed. That is, we seek to learn a function m , mapping from high-dimensional images \mathbf{x}_t to low-dimensional vectors $\mathbf{z}_t \in \mathbb{R}^{n_z}$ with $n_z \ll n_x$, such that the control problem can be solved using \mathbf{z}_t instead of \mathbf{x}_t :

$$\mathbf{z}_t = m(\mathbf{x}_t) + \boldsymbol{\omega}, \quad \boldsymbol{\omega} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\omega}), \quad (2)$$

where $\boldsymbol{\omega}$ accounts for system noise; or equivalently $\mathbf{z}_t \sim \mathcal{N}(m(\mathbf{x}_t), \boldsymbol{\Sigma}_{\omega})$. Assuming for the moment that such a function can be learned (or approximated), we will first define SOC in a latent space and introduce our model thereafter.

2.2 Stochastic locally optimal control in latent spaces

Let $\mathbf{z}_t \in \mathbb{R}^{n_z}$ be the inferred latent state from image \mathbf{x}_t of state \mathbf{s}_t and $f^{\text{lat}}(\mathbf{z}_t, \mathbf{u}_t)$ the transition dynamics in latent space, i.e., $\mathbf{z}_{t+1} = f^{\text{lat}}(\mathbf{z}_t, \mathbf{u}_t)$. Thus f^{lat} models the changes that occur in \mathbf{z}_t when control \mathbf{u}_t is applied to the underlying system as a latent space analogue to $f(\mathbf{s}_t, \mathbf{u}_t)$. Assuming f^{lat} is known, optimal controls for a trajectory of length T in the dynamical system can be derived by minimizing the function $J(\mathbf{z}_{1:T}, \mathbf{u}_{1:T})$ which gives the expected future costs when following $(\mathbf{z}_{1:T}, \mathbf{u}_{1:T})$:

$$J(\mathbf{z}_{1:T}, \mathbf{u}_{1:T}) = \mathbb{E}_{\mathbf{z}} \left[c_T(\mathbf{z}_T, \mathbf{u}_T) + \sum_{t_0}^{T-1} c(\mathbf{z}_t, \mathbf{u}_t) \right], \quad (3)$$

where $c(\mathbf{z}_t, \mathbf{u}_t)$ are instantaneous costs, $c_T(\mathbf{z}_T, \mathbf{u}_T)$ denotes terminal costs and $\mathbf{z}_{1:T} = \{\mathbf{z}_1, \dots, \mathbf{z}_T\}$ and $\mathbf{u}_{1:T} = \{\mathbf{u}_1, \dots, \mathbf{u}_T\}$ are state and action sequences respectively. If \mathbf{z}_t contains sufficient information about \mathbf{s}_t , i.e., \mathbf{s}_t can be inferred from \mathbf{z}_t alone, and f^{lat} is differentiable, the cost-minimizing controls can be computed from $J(\mathbf{z}_{1:T}, \mathbf{u}_{1:T})$ via SOC algorithms [10]. These optimal control algorithms approximate the global non-linear dynamics with locally linear dynamics at each time step t . Locally optimal actions can then be found in closed form. Formally, given a reference trajectory $\bar{\mathbf{z}}_{1:T}$ – the current estimate for the optimal trajectory – together with corresponding controls $\bar{\mathbf{u}}_{1:T}$ the system is linearized as

$$\mathbf{z}_{t+1} = \mathbf{A}(\bar{\mathbf{z}}_t)\mathbf{z}_t + \mathbf{B}(\bar{\mathbf{z}}_t)\mathbf{u}_t + \mathbf{o}(\bar{\mathbf{z}}_t) + \boldsymbol{\omega}, \quad \boldsymbol{\omega} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\omega}), \quad (4)$$

where $\mathbf{A}(\bar{\mathbf{z}}_t) = \frac{\delta f^{\text{lat}}(\bar{\mathbf{z}}_t, \bar{\mathbf{u}}_t)}{\delta \bar{\mathbf{z}}_t}$, $\mathbf{B}(\bar{\mathbf{z}}_t) = \frac{\delta f^{\text{lat}}(\bar{\mathbf{z}}_t, \bar{\mathbf{u}}_t)}{\delta \bar{\mathbf{u}}_t}$ are local Jacobians, and $\mathbf{o}(\bar{\mathbf{z}}_t)$ is an offset. To enable efficient computation of the local controls we assume the costs to be a quadratic function of the latent representation

$$c(\mathbf{z}_t, \mathbf{u}_t) = (\mathbf{z}_t - \mathbf{z}_{\text{goal}})^T \mathbf{R}_z (\mathbf{z}_t - \mathbf{z}_{\text{goal}}) + \mathbf{u}_t^T \mathbf{R}_u \mathbf{u}_t, \quad (5)$$

where $\mathbf{R}_z \in \mathbb{R}^{n_z \times n_z}$ and $\mathbf{R}_u \in \mathbb{R}^{n_u \times n_u}$ are cost weighting matrices and \mathbf{z}_{goal} is the inferred representation of the goal state. We also assume $c_T(\mathbf{z}_T, \mathbf{u}_T) = c(\mathbf{z}_T, \mathbf{u}_T)$ throughout this paper. In combination with Equation (4) this gives us a local *linear-quadratic-Gaussian* formulation at each time step t which can be solved by SOC algorithms such as iterative linear-quadratic regulation (iLQR) [11] or approximate inference control (AICO) [12]. The result of this trajectory optimization step is a locally optimal trajectory with corresponding control sequence $(\mathbf{z}_{1:T}^*, \mathbf{u}_{1:T}^*) \approx \arg \min_{\mathbf{z}_{1:T}, \mathbf{u}_{1:T}} J(\mathbf{z}_{1:T}, \mathbf{u}_{1:T})$.

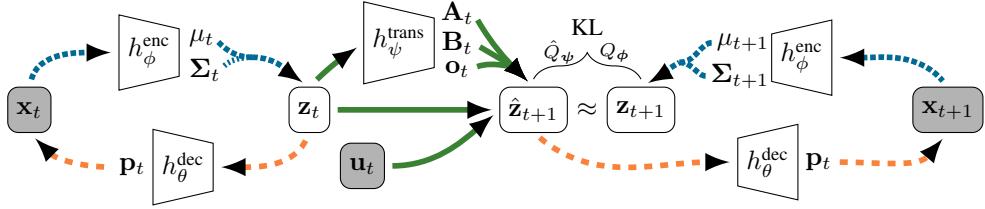


Figure 1: The information flow in the E2C model. From left to right, we encode and decode an image \mathbf{x}_t with the networks h_ϕ^{enc} and h_θ^{dec} , where we use the latent code \mathbf{z}_t for the transition step. The h_ψ^{trans} network computes the local matrices $\mathbf{A}_t, \mathbf{B}_t, \mathbf{o}_t$ with which we can predict $\hat{\mathbf{z}}_{t+1}$ from \mathbf{z}_t and \mathbf{u}_t . Similarity to the encoding \mathbf{z}_{t+1} is enforced by a KL divergence on their distributions and reconstruction is again performed by h_θ^{dec} .

2.3 A locally linear latent state space model for dynamical systems

Starting from the SOC formulation, we now turn to the problem of learning an appropriate low-dimensional latent representation $\mathbf{z}_t \sim P(Z_t|m(\mathbf{x}_t), \Sigma_\omega)$ of \mathbf{x}_t . The representation \mathbf{z}_t has to fulfill three properties: (i) it must capture sufficient information about \mathbf{x}_t (enough to enable reconstruction); (ii) it must allow for accurate prediction of the next latent state \mathbf{z}_{t+1} and thus, implicitly, of the next observation \mathbf{x}_{t+1} ; (iii) the prediction f^{lat} of the next latent state must be locally linearizable *for all valid control magnitudes* \mathbf{u}_t . Given some representation \mathbf{z}_t , properties (ii) and (iii) in particular require us to capture possibly highly non-linear changes of the latent representation due to transformations of the observed scene induced by control commands. Crucially, these are particularly hard to model and subsequently linearize. We circumvent this problem by taking a more direct approach: instead of learning a latent space \mathbf{z} and transition model f^{lat} which are then linearized and combined with SOC algorithms, we directly impose desired transformation properties on the representation \mathbf{z}_t during learning. We will select these properties such that prediction in the latent space as well as locally linear inference of the next observation according to Equation (4) are easy.

The transformation properties that we desire from a latent representation can be formalized directly from the iLQG formulation given in Section 2.2. Formally, following Equation (2), let the latent representation be Gaussian $P(Z|X) = \mathcal{N}(m(\mathbf{x}_t), \Sigma_\omega)$. To infer \mathbf{z}_t from \mathbf{x}_t we first require a method for sampling latent states. Ideally, we would generate samples directly from the unknown true posterior $P(Z|X)$, which we, however, have no access to. Following the variational Bayes approach (see Jordan et al. [13] for an overview) we resort to sampling \mathbf{z}_t from an approximate posterior distribution $Q_\phi(Z|X)$ with parameters ϕ .

Inference model for Q_ϕ . In our work this is always a diagonal Gaussian distribution $Q_\phi(Z|X) = \mathcal{N}(\mu_t, \text{diag}(\sigma_t^2))$, whose mean $\mu_t \in \mathbb{R}^{n_z}$ and covariance $\Sigma_t = \text{diag}(\sigma_t^2) \in \mathbb{R}^{n_z \times n_z}$ are computed by an encoding neural network with outputs

$$\mu_t = \mathbf{W}_\mu h_\phi^{\text{enc}}(\mathbf{x}_t) + \mathbf{b}_\mu, \quad (6)$$

$$\log \sigma_t = \mathbf{W}_\sigma h_\phi^{\text{enc}}(\mathbf{x}_t) + \mathbf{b}_\sigma, \quad (7)$$

where $h_\phi^{\text{enc}} \in \mathbb{R}^{n_e}$ is the activation of the last hidden layer and where ϕ is given by the set of all learnable parameters of the encoding network, including the weight matrices $\mathbf{W}_\mu, \mathbf{W}_\sigma$ and biases $\mathbf{b}_\mu, \mathbf{b}_\sigma$. Parameterizing the mean and variance of a Gaussian distribution based on a neural network gives us a natural and very expressive model for our latent space. It additionally comes with the benefit that we can use the *reparameterization trick* [6, 7] to backpropagate gradients of a loss function based on samples through the latent distribution.

Generative model for P_θ . Using the approximate posterior distribution Q_ϕ we generate observed samples (images) $\tilde{\mathbf{x}}_t$ and $\tilde{\mathbf{x}}_{t+1}$ from latent samples \mathbf{z}_t and \mathbf{z}_{t+1} by enforcing a locally linear relationship in latent space according to Equation (4), yielding the following generative model

$$\begin{aligned} \mathbf{z}_t &\sim Q_\phi(Z|X) = \mathcal{N}(\mu_t, \Sigma_t), \\ \hat{\mathbf{z}}_{t+1} &\sim \hat{Q}_\psi(\hat{Z}|Z, \mathbf{u}) = \mathcal{N}(\mathbf{A}_t \mu_t + \mathbf{B}_t \mathbf{u}_t + \mathbf{o}_t, \mathbf{C}_t), \\ \tilde{\mathbf{x}}_t, \tilde{\mathbf{x}}_{t+1} &\sim P_\theta(X|Z) = \text{Bernoulli}(\mathbf{p}_t), \end{aligned} \quad (8)$$

where \hat{Q}_ψ is the *next latent state* posterior distribution, which exactly follows the linear form required for stochastic optimal control. With $\boldsymbol{\omega}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{H}_t)$ as an estimate of the system noise,

\mathbf{C} can be decomposed as $\mathbf{C}_t = \mathbf{A}_t \Sigma_t \mathbf{A}_t^T + \mathbf{H}_t$. Note that while the transition dynamics in our generative model operates on the inferred latent space, it takes untransformed controls into account. That is, we aim to learn a latent space such that the transition dynamics in \mathbf{z} linearizes the non-linear observed dynamics in \mathbf{x} and is locally linear in the applied controls \mathbf{u} . Reconstruction of an image from \mathbf{z}_t is performed by passing the sample through multiple hidden layers of a decoding neural network which computes the mean \mathbf{p}_t of the generative Bernoulli distribution¹ $P_\theta(X|Z)$ as

$$\mathbf{p}_t = \mathbf{W}_p h_\theta^{\text{dec}}(\mathbf{z}_t) + \mathbf{b}_p, \quad (9)$$

where $h_\theta^{\text{dec}}(\mathbf{z}_t) \in \mathbb{R}^{n_d}$ is the response of the last hidden layer in the decoding network. The set of parameters for the decoding network, including weight matrix \mathbf{W}_p and bias \mathbf{b}_p , then make up the learned generative parameters θ .

Transition model for \hat{Q}_ψ . What remains is to specify how the linearization matrices $\mathbf{A}_t \in \mathbb{R}^{n_z \times n_z}$, $\mathbf{B}_t \in \mathbb{R}^{n_z \times n_u}$ and offset $\mathbf{o}_t \in \mathbb{R}^{n_z}$ are predicted. Following the same approach as for distribution means and covariance matrices, we predict all local transformation parameters from samples \mathbf{z}_t based on the hidden representation $h_\psi^{\text{trans}}(\mathbf{z}_t) \in \mathbb{R}^{n_t}$ of a third neural network with parameters ψ – to which we refer as the transformation network. Specifically, we parametrize the transformation matrices and offset as

$$\begin{aligned} \text{vec}[\mathbf{A}_t] &= \mathbf{W}_A h_\psi^{\text{trans}}(\mathbf{z}_t) + \mathbf{b}_A, \\ \text{vec}[\mathbf{B}_t] &= \mathbf{W}_B h_\psi^{\text{trans}}(\mathbf{z}_t) + \mathbf{b}_B, \\ \mathbf{o}_t &= \mathbf{W}_o h_\psi^{\text{trans}}(\mathbf{z}_t) + \mathbf{b}_o, \end{aligned} \quad (10)$$

where vec denotes vectorization and therefore $\text{vec}[\mathbf{A}_t] \in \mathbb{R}^{(n_z^2)}$ and $\text{vec}[\mathbf{B}_t] \in \mathbb{R}^{(n_z \cdot n_u)}$. To circumvent estimating the full matrix \mathbf{A}_t of size $n_z \times n_z$, we can choose it to be a perturbation of the identity matrix $\mathbf{A}_t = (\mathbf{I} + \mathbf{v}_t \mathbf{r}_t^T)$ which reduces the parameters to be estimated for \mathbf{A}_t to $2n_z$.

A sketch of the complete architecture is shown in Figure 1. It also visualizes an additional constraint that is essential for learning a representation for long-term predictions: we require samples $\hat{\mathbf{z}}_{t+1}$ from the state transition distribution \hat{Q}_ψ to be similar to the encoding of \mathbf{x}_{t+1} through Q_ϕ . While it might seem that just learning a perfect reconstruction of \mathbf{x}_{t+1} from $\hat{\mathbf{z}}_{t+1}$ is enough, we require multi-step predictions for planning in Z which *must correspond* to valid trajectories in the observed space X . Without enforcing similarity between samples from \hat{Q}_ψ and Q_ϕ , following a transition in latent space from \mathbf{z}_t with action \mathbf{u}_t may lead to a point $\hat{\mathbf{z}}_{t+1}$, from which reconstruction of \mathbf{x}_{t+1} is possible, but that is not a valid encoding (i.e. the model will never encode any image as $\hat{\mathbf{z}}_{t+1}$). Executing another action in $\hat{\mathbf{z}}_{t+1}$ then does not result in a valid latent state – since the transition model is conditional on samples coming from the inference network – and thus long-term predictions fail. In a nutshell, such a divergence between encodings and the transition model results in a generative model that does not accurately model the Markov chain formed by the observations.

2.4 Learning via stochastic gradient variational Bayes

For training the model we use a data set $\mathcal{D} = \{(\mathbf{x}_1, \mathbf{u}_1, \mathbf{x}_2), \dots, (\mathbf{x}_{T-1}, \mathbf{u}_{T-1}, \mathbf{x}_T)\}$ containing observation tuples with corresponding controls obtained from interactions with the dynamical system. Using this data set, we learn the parameters of the inference, transition and generative model by minimizing a variational bound on the true data negative log-likelihood $-\log P(\mathbf{x}_t, \mathbf{u}_t, \mathbf{x}_{t+1})$ plus an additional constraint on the latent representation. The complete loss function² is given as

$$\mathcal{L}(\mathcal{D}) = \sum_{(\mathbf{x}_t, \mathbf{u}_t, \mathbf{x}_{t+1}) \in \mathcal{D}} \mathcal{L}^{\text{bound}}(\mathbf{x}_t, \mathbf{u}_t, \mathbf{x}_{t+1}) + \lambda \text{KL}\left(\hat{Q}_\psi(\hat{Z} | \mu_t, \mathbf{u}_t) \| Q_\phi(Z | \mathbf{x}_{t+1})\right). \quad (11)$$

The first part of this loss is the per-example variational bound on the log-likelihood

$$\mathcal{L}^{\text{bound}}(\mathbf{x}_t, \mathbf{u}_t, \mathbf{x}_{t+1}) = \mathbb{E}_{\substack{\mathbf{z}_t \sim Q_\phi \\ \hat{\mathbf{z}}_{t+1} \sim \hat{Q}_\psi}} [-\log P_\theta(\mathbf{x}_t | \mathbf{z}_t) - \log P_\theta(\mathbf{x}_{t+1} | \hat{\mathbf{z}}_{t+1})] + \text{KL}(Q_\phi \| P(Z)), \quad (12)$$

where Q_ϕ , P_θ and \hat{Q}_ψ are the parametric inference, generative and transition distributions from Section 2.3 and $P(Z_t)$ is a prior on the approximate posterior Q_ϕ ; which we always chose to be

¹A Bernoulli distribution for P_θ is a common choice when modeling black-and-white images.

²Note that this is the loss for the latent state space model and distinct from the SOC costs.

an isotropic Gaussian distribution with mean zero and unit variance. The second KL divergence in Equation (11) is an additional contraction term with weight λ , that enforces agreement between the transition and inference models. This term is essential for establishing a Markov chain in latent space that corresponds to the real system dynamics (see Section 2.3 above for an in depth discussion). This KL divergence can also be seen as a prior on the latent transition model. Note that all KL terms can be computed analytically for our model (see supplementary for details).

During training we approximate the expectation in $\mathcal{L}(\mathcal{D})$ via sampling. Specifically, we take one sample \mathbf{z}_t for each input \mathbf{x}_t and transform that sample using Equation (10) to give a valid sample $\hat{\mathbf{z}}_{t+1}$ from \hat{Q}_ψ . We then jointly learn all parameters of our model by minimizing $\mathcal{L}(\mathcal{D})$ using SGD.

3 Experimental Results

We evaluate our model on four visual tasks: an agent in a plane with obstacles, a visual version of the classic inverted pendulum swing-up task, balancing a cart-pole system, and control of a three-link arm with larger images. These are described in detail below.

3.1 Experimental Setup

Model training. We consider two different network types for our model: Standard fully connected neural networks with up to three layers, which work well for moderately sized images, are used for the planar and swing-up experiments; A deep convolutional network for the encoder in combination with an up-convolutional network as the decoder which, in accordance with recent findings from the literature [8, 9], we found to be an adequate model for larger images. Training was performed using Adam [14] throughout all experiments. The training data set \mathcal{D} for all tasks was generated by randomly sampling N state observations and actions with corresponding successor states. For the plane we used $N=3,000$ samples, for the inverted pendulum and cart-pole system we used $N=15,000$ and for the arm $N=30,000$. A complete list of architecture parameters and hyperparameter choices as well as an in-depth explanation of the up-convolutional network are specified in the supplementary material. We will make our code and a video containing controlled trajectories for all systems available under <http://ml.informatik.uni-freiburg.de/research/e2c>.

Model variants. In addition to the Embed to Control (E2C) dynamics model derived above, we also consider two variants: By removing the latent dynamics network h_ψ^{trans} , i.e. setting its output to one in Equation (10) – we obtain a variant in which \mathbf{A}_t , \mathbf{B}_t and \mathbf{o}_t are estimated as globally linear matrices (Global E2C). If we instead replace the transition model with a network estimating the dynamics as a non-linear function \hat{f}^{lat} and only linearize during planning, estimating \mathbf{A}_t , \mathbf{B}_t , \mathbf{o}_t as Jacobians to \hat{f}^{lat} as described in Section 2.2, we obtain a variant with nonlinear latent dynamics.

Baseline models. For a thorough comparison and to exhibit the complicated nature of the tasks, we also test a set of baseline models on the plane and the inverted pendulum task (using the same architecture as the E2C model): a standard variational autoencoder (VAE) and a deep autoencoder (AE) are trained on the autoencoding subtask for visual problems. That is, given a data set \mathcal{D} used for training our model, we remove all actions from the tuples in \mathcal{D} and disregard temporal context between images. After autoencoder training we learn a dynamics model in latent space, approximating f^{lat} from Section 2.2. We also consider a VAE variant with a slowness term on the latent representation – a full description of this variant is given in the supplementary material.

Optimal control algorithms. To perform optimal control in the latent space of different models, we employ two trajectory optimization algorithms: iterative linear quadratic regulation (iLQR) [11] (for the plane and inverted pendulum) and approximate inference control (AICO) [12] (all other experiments). For all VAEs both methods operate on the mean of distributions Q_ϕ and \hat{Q}_ψ . AICO additionally makes use of the local Gaussian covariances Σ_t and \mathbf{C}_t . Except for the experiments on the planar system, control was performed in a model predictive control fashion using the receding horizon scheme introduced in [3]. To obtain closed loop control given an image \mathbf{x}_t , it is first passed through the encoder to obtain the latent state \mathbf{z}_t . A locally optimal trajectory is subsequently found by optimizing $(\mathbf{z}_{t:t+T}^*, \mathbf{u}_{t:t+T}^*) \approx \arg \min_{\mathbf{z}_{t:t+T}, \mathbf{u}_{t:t+T}} J(\mathbf{z}_{t:t+T}, \mathbf{u}_{t:t+T})$ with fixed, small horizon T (with $T = 10$ unless noted otherwise). Controls \mathbf{u}_t^* are applied to the system and a transition to \mathbf{z}_{t+1} is observed (by encoding the next image \mathbf{x}_{t+1}). Then a new control sequence – with horizon

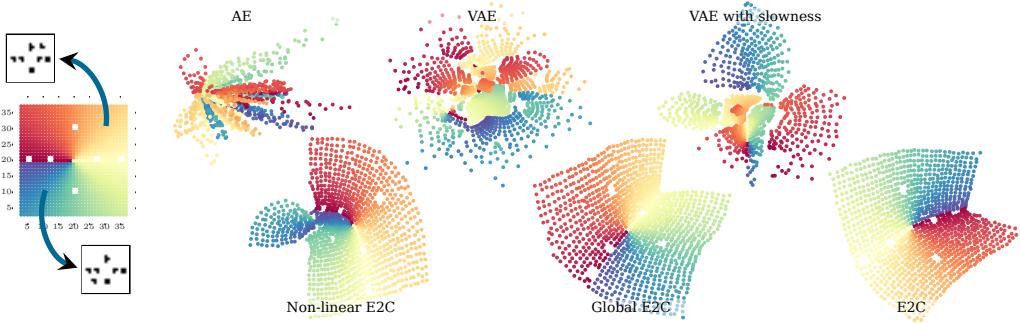


Figure 2: The true state space of the planar system (left) with examples (obstacles encoded as circles) and the inferred spaces (right) of different models. The spaces are spanned by generating images for every valid position of the agent and embedding them with the respective encoders.

$T -$ starting in \mathbf{z}_{t+1} is found using the last estimated trajectory as a bootstrap. Note that planning is performed entirely in the latent state *without access to any observations* except for the depiction of the current state. To compute the cost function $c(\mathbf{z}_t, \mathbf{u}_t)$ required for trajectory optimization in \mathbf{z} we assume knowledge of the observation \mathbf{x}_{goal} of the goal state \mathbf{s}_{goal} . This observation is then transformed into latent space and costs are computed according to Equation (5).

3.2 Control in a planar system

The agent in the planar system can move in a bounded two-dimensional plane by choosing a continuous offset in x- and y-direction. The high-dimensional representation of a state is a 40×40 black-and-white image. Obstructed by six circular obstacles, the task is to move to the bottom right of the image, starting from a random x position at the top of the image. The encodings of obstacles are obtained prior to planning and an additional quadratic cost term is penalizing proximity to them.

A depiction of the observations on which control is performed – together with their corresponding state values and embeddings into latent space – is shown in Figure 2. The figure also clearly shows a fundamental advantage the E2C model has over its competitors: While the separately trained autoencoders make for aesthetically pleasing pictures, the models failed to discover the underlying structure of the state space, complicating dynamics estimation and largely invalidating costs based on distances in said space. Including the latent dynamics constraints in these end-to-end models on the other hand, yields latent spaces approaching the optimal planar embedding.

We test the long-term accuracy by accumulating latent and real trajectory costs to quantify whether the imagined trajectory reflects reality. The results for all models when starting from random positions at the top and executing 40 pre-computed actions are summarized in Table 1 – using a separate test set for evaluating reconstructions. While all methods achieve a low reconstruction loss, the difference in accumulated real costs per trajectory show the superiority of the E2C model. Using the globally or locally linear E2C model, trajectories planned in latent space are as good as trajectories planned on the real state. All models besides E2C fail to give long-term predictions that result in good performance.

3.3 Learning swing-up for an inverted pendulum

We next turn to the task of controlling the classical inverted pendulum system [15] from images. We create depictions of the state by rendering a fixed length line starting from the center of the image at an angle corresponding to the pendulum position. The goal in this task is to swing-up and balance an underactuated pendulum from a resting position (pendulum hanging down). Exemplary observations and reconstructions for this system are given in Figure 3(d). In the visual inverted pendulum task our algorithm faces two additional difficulties: the observed space is non-Markov, as the angular velocity cannot be inferred from a single image, and second, discretization errors due to rendering pendulum angles as small 48×48 pixel images make exact control difficult. To restore the Markov property, we stack two images (as input channels), thus observing a one-step history.

Figure 3 shows the topology of the latent space for our model, as well as one sample trajectory in true state and latent space. The fact that the model can learn a meaningful embedding, separating

Table 1: Comparison between different approaches to model learning from raw pixels for the planar and pendulum system. We compare all models with respect to their prediction quality on a test set of sampled transitions and with respect to their performance when combined with SOC (trajectory cost for control from different start states). Note that trajectory costs in latent space are not necessarily comparable. The “real” trajectory cost was computed on the dynamics of the simulator while executing planned actions. For the true models for s_t , real trajectory costs were 20.24 ± 4.15 for the planar system, and 9.8 ± 2.4 for the pendulum. Success was defined as reaching the goal state and staying ϵ -close to it for the rest of the trajectory (if non terminating). All statistics quantify over 5/30 (plane/pendulum) different starting positions. A \dagger marks separately trained dynamics networks.

Algorithm	State Loss $\log p(x_t \hat{x}_t)$	Next State Loss $\log p(x_{t+1} \hat{x}_t, u_t)$	Trajectory Cost Latent	Trajectory Cost Real	Success percent
Planar System					
AE †	11.5 ± 97.8	3538.9 ± 1395.2	1325.6 ± 81.2	273.3 ± 16.4	0 %
VAE †	3.6 ± 18.9	652.1 ± 930.6	43.1 ± 20.8	91.3 ± 16.4	0 %
VAE + slowness †	10.5 ± 22.8	104.3 ± 235.8	47.1 ± 20.5	89.1 ± 16.4	0 %
Non-linear E2C	8.3 ± 5.5	11.3 ± 10.1	19.8 ± 9.8	42.3 ± 16.4	96.6 %
Global E2C	6.9 ± 3.2	9.3 ± 4.6	12.5 ± 3.9	27.3 ± 9.7	100 %
E2C	7.7 ± 2.0	9.7 ± 3.2	10.3 ± 2.8	25.1 ± 5.3	100 %
Inverted Pendulum Swing-Up					
AE †	8.9 ± 100.3	13433.8 ± 6238.8	1285.9 ± 355.8	194.7 ± 44.8	0 %
VAE †	7.5 ± 47.7	8791.2 ± 17356.9	497.8 ± 129.4	237.2 ± 41.2	0 %
VAE + slowness †	26.5 ± 18.0	779.7 ± 633.3	419.5 ± 85.8	188.2 ± 43.6	0 %
E2C no latent KL	64.4 ± 32.8	87.7 ± 64.2	489.1 ± 87.5	213.2 ± 84.3	0 %
Non-linear E2C	59.6 ± 25.2	72.6 ± 34.5	313.3 ± 65.7	37.4 ± 12.4	63.33 %
Global E2C	115.5 ± 56.9	125.3 ± 62.6	628.1 ± 45.9	125.1 ± 10.7	0 %
E2C	84.0 ± 50.8	89.3 ± 42.9	275.0 ± 16.6	15.4 ± 3.4	90 %

velocities and positions, from this data is remarkable (no other model recovered this shape). Table 1 again compares the different models quantitatively. While the E2C model is not the best in terms of reconstruction performance, it is the only model resulting in stable swing-up *and* balance behavior. We explain the failure of the other models with the fact that the non-linear latent dynamics model cannot be guaranteed to be linearizable for all control magnitudes, resulting in undesired behavior around unstable fixpoints of the real system dynamics, and that for this task a globally linear dynamics model is inadequate.

3.4 Balancing a cart-pole and controlling a simulated robot arm

Finally, we consider control of two more complex dynamical systems from images using a six layer convolutional inference and six layer up-convolutional generative network, resulting in a *12-layer deep* path from input to reconstruction. Specifically, we control a visual version of the classical cart-pole system [16] from a history of two 80×80 pixel images as well as a three-link planar robot arm based on a history of two 128×128 pixel images. The latent space was set to be 8-dimensional in both experiments. The real state dimensionality for the cart-pole is four and is controlled using one

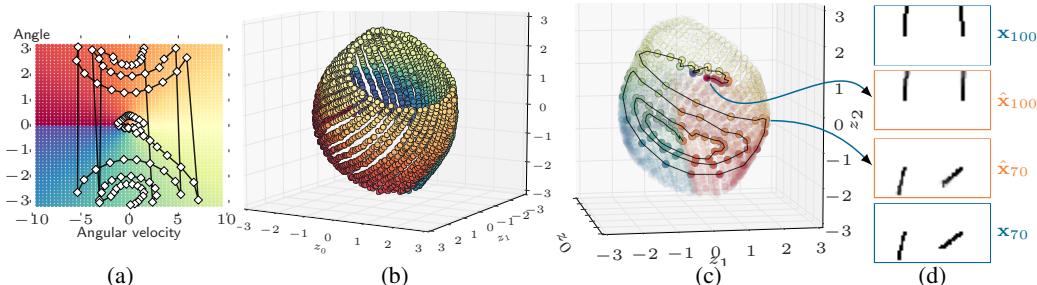


Figure 3: (a) The true state space of the inverted pendulum task overlaid with a successful trajectory taken by the E2C agent. (b) The learned latent space. (c) The trajectory from (a) traced out in the latent space. (d) Images x and reconstructions \hat{x} showing current positions (right) and history (left).

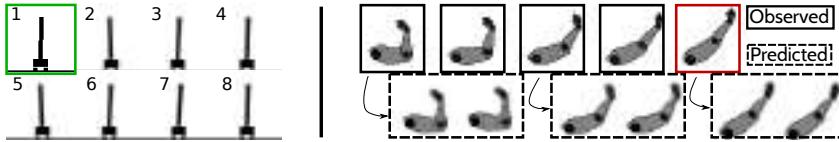


Figure 4: Left: Trajectory from the cart-pole domain. Only the first image (green) is “real”, all other images are “dreamed up” by our model. Notice discretization artifacts present in the real image. Right: Exemplary observed (with history image omitted) and predicted images (including the history image) for a trajectory in the visual robot arm domain with the goal marked in red.

action, while for the arm the real state can be described in 6 dimensions (joint angles and velocities) and controlled using a three-dimensional action vector corresponding to motor torques.

As in previous experiments the E2C model seems to have no problem finding a locally linear embedding of images into latent space in which control can be performed. Figure 4 depicts exemplary images – for both problems – from a trajectory executed by our system. The costs for these trajectories (11.13 for the cart-pole, 85.12 for the arm) are only slightly worse than trajectories obtained by AICO operating on the real system dynamics starting from the same start-state (7.28 and 60.74 respectively). The supplementary material contains additional experiments using these domains.

4 Comparison to recent work

In the context of representation learning for control (see Böhmer et al. [17] for a review), deep autoencoders (ignoring state transitions) similar to our baseline models have been applied previously, e.g. by Lange and Riedmiller [18]. A more direct route to control based on image streams is taken by recent work on (model free) deep end-to-end Q-learning for Atari games by Mnih et al. [19], as well as kernel based [20] and deep policy learning for robot control [21].

Close to our approach is a recent paper by Wahlström et al. [22], where autoencoders are used to extract a latent representation for control from images, on which a non-linear model of the forward dynamics is learned. Their model is trained jointly and is thus similar to the non-linear E2C variant in our comparison. In contrast to our model, their formulation requires PCA pre-processing and does neither ensure that long-term predictions in latent space do not diverge, nor that they are linearizable.

As stated above, our system belongs to the family of VAEs and is generally similar to recent work such as Kingma and Welling [6], Rezende et al. [7], Gregor et al. [23], Bayer and Osendorfer [24]. Two additional parallels between our work and recent advances for training deep neural networks can be observed. First, the idea of enforcing desired transformations in latent space during learning – such that the data becomes easy to model – has appeared several times already in the literature. This includes the development of transforming auto-encoders [25] and recent probabilistic models for images [26, 27]. Second, learning relations between pairs of images – although *without control* – has received considerable attention from the community during the last years [28, 29]. In a broader context our model is related to work on state estimation in Markov decision processes (see Langford et al. [30] for a discussion) through, e.g., hidden Markov models and Kalman filters [31, 32].

5 Conclusion

We presented Embed to Control (E2C), a system for stochastic optimal control on high-dimensional image streams. Key to the approach is the extraction of a latent dynamics model which is constrained to be locally linear in its state transitions. An evaluation on four challenging benchmarks revealed that E2C can find embeddings on which control can be performed with ease, reaching performance close to that achievable by optimal control on the real system model.

Acknowledgments

We thank A. Radford, L. Metz, and T. DeWolf for sharing code, as well as A. Dosovitskiy for useful discussions. This work was partly funded by a DFG grant within the priority program “Autonomous learning” (SPP1597) and the BrainLinks-BrainTools Cluster of Excellence (grant number EXC 1086). M. Watter is funded through the State Graduate Funding Program of Baden-Württemberg.

References

- [1] D. Jacobson and D. Mayne. Differential dynamic programming. *American Elsevier*, 1970.
- [2] E. Todorov and W. Li. A generalized iterative LQG method for locally-optimal feedback control of constrained nonlinear stochastic systems. In *ACC. IEEE*, 2005.
- [3] Y. Tassa, T. Erez, and W. D. Smart. Receding horizon differential dynamic programming. In *Proc. of NIPS*, 2008.
- [4] Y. Pan and E. Theodorou. Probabilistic differential dynamic programming. In *Proc. of NIPS*, 2014.
- [5] S. Levine and V. Koltun. Variational policy search via trajectory optimization. In *Proc. of NIPS*, 2013.
- [6] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *Proc. of ICLR*, 2014.
- [7] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proc. of ICML*, 2014.
- [8] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus. Deconvolutional networks. In *CVPR*, 2010.
- [9] A. Dosovitskiy, J. T. Springenberg, and T. Brox. Learning to generate chairs with convolutional neural networks. In *Proc. of CVPR*, 2015.
- [10] R. F. Stengel. *Optimal Control and Estimation*. Dover Publications, 1994.
- [11] W. Li and E. Todorov. Iterative Linear Quadratic Regulator Design for Nonlinear Biological Movement Systems. In *Proc. of ICINCO*, 2004.
- [12] M. Toussaint. Robot Trajectory Optimization using Approximate Inference. In *Proc. of ICML*, 2009.
- [13] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. In *Machine Learning*, 1999.
- [14] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proc. of ICLR*, 2015.
- [15] H. Wang, K. Tanaka, and M. Griffin. An approach to fuzzy control of nonlinear systems; stability and design issues. *IEEE Trans. on Fuzzy Systems*, 4(1), 1996.
- [16] R. S. Sutton and A. G. Barto. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1st edition, 1998. ISBN 0262193981.
- [17] W. Böhmer, J. T. Springenberg, J. Boedecker, M. Riedmiller, and K. Obermayer. Autonomous learning of state representations for control. *KI - Künstliche Intelligenz*, 2015.
- [18] S. Lange and M. Riedmiller. Deep auto-encoder neural networks in reinforcement learning. In *Proc. of IJCNN*, 2010.
- [19] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540), 02 2015.
- [20] H. van Hoof, J. Peters, and G. Neumann. Learning of non-parametric control policies with high-dimensional state features. In *Proc. of AISTATS*, 2015.
- [21] S. Levine, C. Finn, T. Darrell, and P. Abbeel. End-to-end training of deep visuomotor policies. *CoRR*, abs/1504.00702, 2015. URL <http://arxiv.org/abs/1504.00702>.
- [22] N. Wahlström, T. B. Schön, and M. P. Deisenroth. From pixels to torques: Policy learning with deep dynamical models. *CoRR*, abs/1502.02251, 2015. URL <http://arxiv.org/abs/1502.02251>.
- [23] K. Gregor, I. Danihelka, A. Graves, D. Rezende, and D. Wierstra. DRAW: A recurrent neural network for image generation. In *Proc. of ICML*, 2015.
- [24] J. Bayer and C. Osendorfer. Learning stochastic recurrent networks. In *NIPS 2014 Workshop on Advances in Variational Inference*, 2014.
- [25] G. Hinton, A. Krizhevsky, and S. Wang. Transforming auto-encoders. In *Proc. of ICANN*, 2011.
- [26] L. Dinh, D. Krueger, and Y. Bengio. Nice: Non-linear independent components estimation. *CoRR*, abs/1410.8516, 2015. URL <http://arxiv.org/abs/1410.8516>.
- [27] T. Cohen and M. Welling. Transformation properties of learned visual representations. In *ICLR*, 2015.
- [28] G. W. Taylor, L. Sigal, D. J. Fleet, and G. E. Hinton. Dynamical binary latent variable models for 3d human pose tracking. In *Proc. of CVPR*, 2010.
- [29] R. Memisevic. Learning to relate images. *IEEE Trans. on PAMI*, 35(8):1829–1846, 2013.
- [30] J. Langford, R. Salakhutdinov, and T. Zhang. Learning nonlinear dynamic models. In *ICML*, 2009.
- [31] M. West and J. Harrison. *Bayesian Forecasting and Dynamic Models (Springer Series in Statistics)*. Springer-Verlag, February 1997. ISBN 0387947256.

- [32] T. Matsubara, V. Gómez, and H. J. Kappen. Latent Kullback Leibler control for continuous-state systems using probabilistic graphical models. *UAI*, 2014.
- [33] T. D. Kulkarni, W. Whitney, P. Kohli, and J. B. Tenenbaum. Deep convolutional inverse graphics network. *CoRR*, abs/1503.03167, 2015. URL <http://arxiv.org/abs/1503.03167>.
- [34] C. Osendorfer, H. Soyer, and P. van der Smagt. Image super-resolution with fast approximate convolutional sparse coding. In *Proc. of ICONIP*, Lecture Notes in Computer Science. Springer International Publishing, 2014.
- [35] R. Jonschkowski and O. Brock. State representation learning in robotics: Using prior knowledge about physical interaction. In *Proc. of RSS*, 2014.
- [36] R. Legenstein, N. Wilbert, and L. Wiskott. Reinforcement learning on slow features of high-dimensional input streams. *PLoS Computational Biology*, 2010.
- [37] W. Zou, A. Ng, and K. Yu. Unsupervised learning of visual invariance with temporal coherence. In *NIPS*2011 Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.
- [38] A. M. Saxe, J. L. McClelland, and S. Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *Proc. of ICLR*, 2014.
- [39] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In *AISTATS*. Journal of Machine Learning Research - Workshop and Conference Proceedings, 2011.

A Supplementary to the E2C description

A.1 State transition matrix factorization and KL Divergence

As alluded to in the main paper, estimation of the full local state transition matrix $\mathbf{A}_t \in \mathbb{R}^{n_z \times n_z}$ from Equation (8) requires the transition network to predict $n_z \times n_z$ parameters. Using an arbitrary state transition matrix also – inconveniently – requires inversion of said matrix for computing the KL divergence penalty from Equation (11) (through which it is hard to backpropagate). We started our experiments using a full matrix (and only approximating all KL divergence terms), but quickly found that a rank one perturbation of the identity matrix could be used instead without loss of performance in any of our benchmarks. To the contrary, the resulting networks have fewer parameters and are thus easier to train. We here give the derivation of this process and how the KL divergence from Equation (11) can be computed. For the reformulation we represent \mathbf{A}_t as $\mathbf{A}_t = \mathbf{I} + \mathbf{v}_t \mathbf{r}_t^T$, therefore only \mathbf{v}_t and \mathbf{r}_t need to be estimated by the transition network, reducing the number of outputs for \mathbf{A}_t from n_z^2 to $2n_z$.

The KL divergence between two multivariate Gaussians is given by

$$\text{KL}(\mathcal{N}_0 || \mathcal{N}_1) = \frac{1}{2} \left(\text{Tr} (\Sigma_1^{-1} \Sigma_0) + (\mu_1 - \mu_0)^T \Sigma_1^{-1} (\mu_1 - \mu_0) - k + \log \left(\frac{\det \Sigma_1}{\det \Sigma_0} \right) \right). \quad (13)$$

For a simplified notation, such that $\text{KL}(\mathcal{N}_0 || \mathcal{N}_1) = \text{KL}(\hat{Q} || Q)$, let us assume

$$\begin{aligned} \mathcal{N}_0 &= \mathcal{N}(\mu_0, \mathbf{A} \Sigma_0 \mathbf{A}^T) = \mathcal{N}(\mu_t, \mathbf{A}_t \Sigma_t \mathbf{A}_t^T) = \hat{Q}, \\ \mathcal{N}_1 &= \mathcal{N}(\mu_1, \Sigma_1) = \mathcal{N}(\mu_{t+1}, \Sigma_{t+1}) = Q. \end{aligned}$$

The main point behind the derivation presented in the following, is to make partial derivatives of the above KL divergence efficiently computable. To this end, we cannot take the trace or the determinant via numerical algorithms, because we have to be able to take the gradients in symbolic form. Aside from that, we like to process a batch of samples, so the computation should have a convenient form and not require excessive amounts of tensor products in between. We start our simplification with

the trace term which results in

$$\begin{aligned}
\text{Tr}(\Sigma_1^{-1}\Sigma_0) &= \text{Tr}(\Sigma_1^{-1}\mathbf{A}\Sigma_0\mathbf{A}^T) \\
&= \text{Tr}(\Sigma_1^{-1}(\mathbf{I}+\mathbf{v}\mathbf{r}^T)\Sigma_0(\mathbf{I}+\mathbf{v}\mathbf{r}^T)^T) \\
&= \text{Tr}((\Sigma_1^{-1} + \Sigma_1^{-1}\mathbf{v}\mathbf{r}^T)(\Sigma_0 + \Sigma_0(\mathbf{v}\mathbf{r}^T)^T)) \\
&= \text{Tr}(\Sigma_1^{-1}\Sigma_0 + \Sigma_1^{-1}\Sigma_0(\mathbf{v}\mathbf{r}^T)^T + \Sigma_1^{-1}\mathbf{v}\mathbf{r}^T\Sigma_0 + \Sigma_1^{-1}\mathbf{v}\mathbf{r}^T\Sigma_0(\mathbf{v}\mathbf{r}^T)^T) \\
&\quad \text{Tr}(A+B) = \text{Tr}(A) + \text{Tr}(B) \\
&= \text{Tr}(\Sigma_1^{-1}\Sigma_0) + \text{Tr}(\Sigma_1^{-1}\Sigma_0(\mathbf{v}\mathbf{r}^T)^T) + \text{Tr}(\Sigma_1^{-1}\mathbf{v}\mathbf{r}^T\Sigma_0) + \text{Tr}(\Sigma_1^{-1}\mathbf{v}\mathbf{r}^T\Sigma_0\mathbf{r}\mathbf{v}^T) \\
&\quad \text{Tr}(ABC) = \text{Tr}(CAB) = \dots \\
&= \sum_i \frac{\sigma_{0,i}^2}{\sigma_{1,i}^2} + \sum_i \frac{\sigma_{0,i}^2 r_i v_i}{\sigma_{1,i}^2} + \sum_i \frac{v_i r_i \sigma_{0,i}^2}{\sigma_{1,i}^2} + \text{Tr}(\mathbf{v}^T \Sigma_1^{-1} \mathbf{v} \mathbf{r}^T \Sigma_0 \mathbf{r}) \\
&= \sum_i \frac{\sigma_{0,i}^2 + 2\sigma_{0,i}^2 v_i r_i}{\sigma_{1,i}^2} + \sum_i r_i^2 \sigma_i^2 \cdot \sum_i \frac{v_i^2}{\sigma_i^2}.
\end{aligned}$$

The last equation is easy to implement and only requires summing over the non-batch dimension. The difference of means can be derived very quickly with the same summing scheme:

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \Sigma_1^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) = \sum_i \frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)_i^2}{\sigma_{1,i}^2}.$$

It remains the ratio of determinants, which we will simplify with the matrix determinant lemma giving

$$\begin{aligned}
\log \left(\frac{\det \Sigma_1}{\det \mathbf{A} \Sigma_0 \mathbf{A}^T} \right) &= \log \det \Sigma_1 - \log \det (\mathbf{A} \Sigma_0 \mathbf{A}^T) \\
&= \log \prod_i \sigma_{1,i}^2 - \log (\det \mathbf{A} \cdot \det \Sigma_0 \cdot \det \mathbf{A}^T) \quad \text{det } \mathbf{A}^T = \det \mathbf{A} \\
&= 2 \sum_i \log \sigma_{1,i} - \log \left((\det \mathbf{A})^2 \prod_i \sigma_{0,i}^2 \right) \quad \text{Matrix determinant lemma} \\
&= 2 \sum_i \log \sigma_{1,i} - \log (1 + \mathbf{v}^T \mathbf{r})^2 - 2 \sum_i \log \sigma_{0,i} \\
&= 2 \left(\sum_i (\log \sigma_{1,i}^2 - \log \sigma_{0,i}^2) - \log(1 + \sum_i v_i r_i) \right).
\end{aligned}$$

Putting the above two formulas together finally yields

$$\begin{aligned}
\text{KL}(\mathcal{N}_0 || \mathcal{N}_1) &= \frac{1}{2} \left(\sum_i \frac{\sigma_{0,i}^2 + 2\sigma_{0,i}^2 v_i r_i}{\sigma_{1,i}^2} + \sum_i r_i^2 \sigma_i^2 \cdot \sum_i \frac{v_i^2}{\sigma_i^2} \right. \\
&\quad \left. + \sum_i \frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)_i^2}{\sigma_{1,i}^2} - k \right. \\
&\quad \left. + 2 \left(\sum_i (\log \sigma_{1,i}^2 - \log \sigma_{0,i}^2) - \log(1 + \sum_i v_i r_i) \right) \right). \tag{14}
\end{aligned}$$

B Supplementary to the experimental setup

B.1 Up-convolution

We used convolutional inference networks for the cart-pole and three-link arm task. While these networks help us overcome the problem of large input dimensionalities (i.e. $2 \times 128 \times 128$ pixel

images in the three-link arm task), we still have to generate full resolution images with the decoder network. For high-dimensional images generation fully connected neural networks are simply not an option. We thus decided to use up-convolutional networks, which were recently shown to be powerful models for image generation [8, 9, 33].

To set-up these models we basically “mirror” the convolutional architecture used for the encoder. More specifically for each 5×5 convolution followed by 2×2 max-pooling step in the encoder network, we introduce a 2×2 up-sampling and 5×5 convolution step in the decoder network. The complete network architecture is given below. It is similar to the up-convolution networks used in Dosovitskiy et al. [9]. The upsampling strategy we use is simple “perforated” upsampling as described in [34].

B.2 Variational Autoencoder with slowness

Enforcing temporal slowness during learning has previously been found to be a good proxy for learning representations in reinforcement learning [35, 36] and representation learning from videos [37]. We also consider a VAE variant with a slowness term on the latent representation by enforcing similarity of the encodings of temporally close images. This can be achieved by augmenting the standard VAE objective $\mathcal{L}^{\text{bound}}$ with an additional KL divergence term on the latent posterior Q_ϕ :

$$\mathcal{L}^{\text{slow}}(\mathbf{x}_t, \mathbf{x}_{t+1}) = \text{KL}(Q_\phi(\mathbf{z}_{t+1}|\mathbf{x}_{t+1})\|Q_\phi(\mathbf{z}_t|\mathbf{x}_t)). \quad (15)$$

Indeed there seems to be a slightly better coherence of similar states in the latent spaces, as e.g. depicted in Figure 8 in the main paper. Yet, our experiments show that a slowness term alone does not suffice to structure the latent space, such that locally linear predictions and control become feasible.

B.3 Evaluation criteria

For comparing the performance of all variants of E2C and the baselines, the following criteria are of importance:

- **Autoencoding.** Being able to reconstruct the given observations is the basic necessity for a model to work. The reconstruction cost drives a model to identify single states from its observations.
- **Decoding the next state.** For any planning to be possible at all, the decoder must be able to generate the correct images from transitions the dynamics model performed. If this is not the case, we know that the latent states of the encoding and the transition model do not coincide, thus preventing any planning.
- **Optimizing latent trajectory costs.** The action sequences for achieving a specified goal will be determined completely by locally linearized dynamics in the latent space. Therefore minimizing trajectory costs in latent space is, again, a necessity for successful control.
- **Optimizing real trajectory costs.** While the action sequence has been determined for the latent dynamics, the deciding criterion is whether this reflects the true state trajectory costs. Therefore carrying out the “dreamed” plans in reality is the optimality criterion for every model. To make the different models comparable, we use the same cost matrices for evaluation, which are not necessarily the same as for optimization.

We reflected these four criteria in the evaluation table in the paper. For the reconstruction of the current and next state we specified the mean log loss, which is in case of the Bernoulli distributions the cross entropy error function:

$$\log p(\mathbf{x}|\hat{\mathbf{x}}) = \frac{1}{N} \sum_{n=1}^N \sum_{i=0}^{n_x} x_{n,i} \log \hat{x}_{n,i} + (1 - x_{n,i}) \log(1 - \hat{x}_{n,i}). \quad (16)$$

For the costs a model imagines and truly achieves, we sample from different starting states and accumulate the distances in latent and true state space according to the SOC method.

B.4 The three-link robot arm

The robot arm we used in the last experiment in the main paper was simulated using dynamics generated by the MapleSim <http://www.maplesoft.com/products/maplesim/> simulator wrapped in Python and visualized for producing inputs to E2C using PyGame. We simulated a fairly standard robot arm with three links. The length of the links were set to 2, 1.2 and 0.7 (units were set to meters). The masses of the corresponding links were all set to 10kg .

B.5 Evaluating the true system model

To compare the efficacy of different models when combined with optimal control algorithms, we always reported the cost in latent space (as used by the optimal control algorithm) as well as the “real” trajectory cost. To compute this real cost, we evaluated the same cost function as in the latent space (quadratic costs on the deviation from a given goal state), but using the real system states during execution and different cost matrices for a fair comparison.

As an upper bound on the performance achievable for control by any of the models, we also computed the true system cost by applying iLQR/AICO to a model of the real system dynamics. We have this model available since all experiments were performed in simulation.

B.6 Neural Network training

B.6.1 Experimental Setup

All the datasets were created in advance as $\mathcal{D} = \{(\mathbf{x}_1, \mathbf{u}_1, \mathbf{x}_2), \dots, (\mathbf{x}_{T-1}, \mathbf{u}_{T-1}, \mathbf{x}_T)\}$ for the training, validation and test split. While the E2C models were trained on \mathcal{D} , the ones that do not incorporate any transition information (i.e. AE, VAE) were trained on images $\mathcal{D}_{\text{images}} = \{\mathbf{x}_1, \dots, \mathbf{x}_T\}$ extracted from the original dataset \mathcal{D} . The slowness VAE was trained on the pairs of images subset $\mathcal{D}_{\text{pairs}} = \{(\mathbf{x}_1, \mathbf{x}_2), \dots, (\mathbf{x}_{T-1}, \mathbf{x}_T)\}$ and our E2C models on the full \mathcal{D} .

In order to learn dynamics predictions for the image-only autoencoders, we extracted the latent representations and combined them with the actions from \mathcal{D} into $\mathcal{D}_{\text{dynamics}} = \{(\mathbf{z}_1, \mathbf{u}_1, \mathbf{z}_2), \dots, (\mathbf{z}_{T-1}, \mathbf{u}_{T-1}, \mathbf{z}_T)\}$. On these low-dimensional representations we trained the dynamics MLPs, thus ensuring that all methods were trained on exactly the same data.

B.6.2 Implementation details

We used orthogonal weight initialization for every layer [38]. As described in the main paper, Adam [14] was used as the learning rule for all networks. We found both these techniques to be fundamentally important for stabilizing training and achieving good reconstructions for all methods. Both methods also clearly helped to cut the hyperparameter search needed for all methods to a minimum. In the process of training, we could make out three phases: the unfolding of the latent space, the overcoming of the trivial solution (the average image of the dataset) and the minimization of the latent KL term. The architectures used for our experiments were as follows (where ReLU stands for rectified linear units [39] and conv. for convolutions):

Plane

- Input: 40^2 image dimensions, 2 action dimensions
- Latent Space dimensionality: 2
- Encoder: 150 ReLU - 150 ReLU - 150 ReLU - 4 Linear (2 for AE)
- Decoder: 200 ReLU - 200 ReLU - 1600 Linear (Sigmoid for AE)
- Dynamics: 100 ReLU - 100 ReLU + Output layer (except Global E2C)
 - AE, VAE with slowness, Non-linear E2C: 2 Linear
 - E2C: 8 Linear (2 · 2 for \mathbf{A}_t , 2 · 1 for \mathbf{B}_t , 2 for \mathbf{o}_t), $\lambda = 0.25$
- Adam: $\alpha = 10^{-4}$, $\beta_2 = 0.1$
- Evaluation costs: $\mathbf{R}_z = 0.1 \cdot \mathbf{I}$, $\mathbf{R}_u = \mathbf{I}$, $\mathbf{R}_o = \mathbf{I}$

Pendulum swing-up

- Input: $2 \cdot 48^2$ image dimensions, 1 action dimension
- Latent Space dimensionality: 3
- Encoder: 800 ReLU - 800 ReLU - 6 Linear (3 for AE)
- Decoder: 800 ReLU - 800 ReLU - 4608 Linear (Sigmoid for AE)
- Dynamics: 100 ReLU - 100 ReLU + Output layer (except Global E2C)
 - AE, VAE, VAE with slowness, Non-linear E2C: 3 Linear
 - E2C: 12 Linear ($2 \cdot 3$ for $\mathbf{A}_t = (\mathbf{I} + \mathbf{v}_t \mathbf{r}_t^T)$, $3 \cdot 1$ for \mathbf{B}_t , 3 for \mathbf{b}_t), $\lambda = 0.25$
- Adam: $\alpha = 3 \cdot 10^{-4}$, $\beta_2 = 0.1$
- Evaluation costs: $\mathbf{R}_z = \mathbf{I}$, $\mathbf{R}_u = 0.1\mathbf{I}$

Cart-Pole balancing

- Input: $2 \cdot 80^2$ image dimensions, 1 action dimension
- Latent Space dimensionality: 8
- Encoder: $32 \times 5 \times 5$ ReLU - $32 \times 5 \times 5$ ReLU - $32 \times 5 \times 5$ ReLU - 512 ReLU - 512 ReLU
- Decoder: 512 ReLU - 512 ReLU - 2×2 up-sampling - $32 \times 5 \times 5$ ReLU - 2×2 up-sampling - $32 \times 5 \times 5$ ReLU - 2×2 up-sampling - $32 \times 5 \times 5$ conv. ReLU
- Dynamics: 200 ReLU - 200 ReLU + 32 Linear ($2 \cdot 8$ for $\mathbf{A}_t = (\mathbf{I} + \mathbf{v}_t \mathbf{r}_t^T)$, $8 \cdot 1$ for \mathbf{B}_t , 8 for \mathbf{b}_t), $\lambda = 1$
- Adam: $\alpha = 10^{-4}$, $\beta_2 = 0.1$
- Evaluation costs: $\mathbf{R}_z = \mathbf{I}$, $\mathbf{R}_u = \mathbf{I}$

Three-link arm

- Input: $2 \cdot 128^2$ image dimensions, 3 action dimensions
- Latent Space dimensionality: 8
- Encoder: $64 \times 5 \times 5$ conv. ReLU - 2×2 max-pooling - $32 \times 5 \times 5$ conv. ReLU - 2×2 max-pooling - $32 \times 5 \times 5$ conv. ReLU - 2×2 max-pooling - 512 ReLU - 512 ReLU
- Decoder: 512 ReLU - 512 ReLU - 2×2 up-sampling - $32 \times 5 \times 5$ ReLU - 2×2 up-sampling - $32 \times 5 \times 5$ ReLU - 2×2 up-sampling - $64 \times 5 \times 5$ conv. ReLU
- Dynamics: 200 ReLU - 200 ReLU + 48 Linear ($2 \cdot 8$ for $\mathbf{A}_t = (\mathbf{I} + \mathbf{v}_t \mathbf{r}_t^T)$, $8 \cdot 3$ for \mathbf{B}_t , 8 for \mathbf{b}_t), $\lambda = 1$
- Adam: $\alpha = 10^{-4}$, $\beta_2 = 0.1$
- Evaluation costs: $\mathbf{R}_z = \mathbf{I}$, $\mathbf{R}_u = 0.001\mathbf{I}$

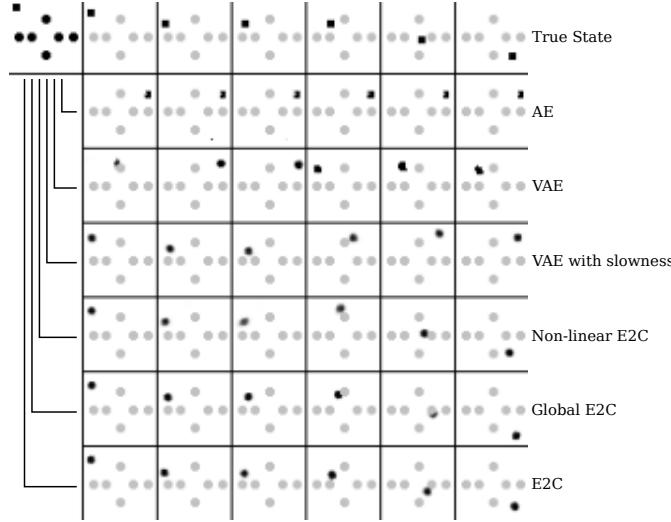


Figure 5: Generated “dreamed” trajectories of different models for the plane task (from left to right). The opacity of the obstacles has been lowered in this depiction for better visibility of the agent.

C Supplementary evaluations

C.1 Trajectories for plane and pendulum

To qualitatively measure the predictive accuracy, the starting state for a trajectory is encoded and the actions are applied on the latent representation. After each transition, the predicted latent position is decoded and visualized. In this manner, multi-step predictions can be generated for the planar system in Figure 5 and for the inverted pendulum in Figures 6 and 7.

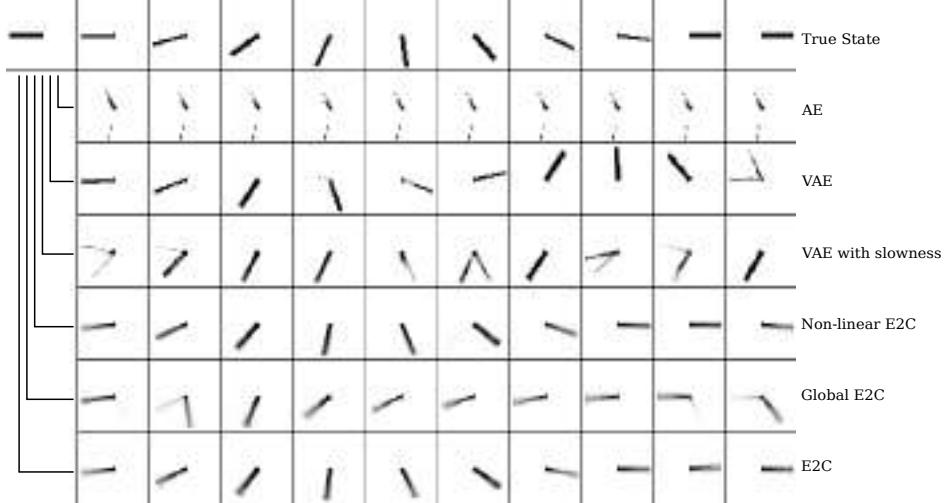


Figure 6: Generated “dreamed” trajectories (from left to right) for *passive* dynamics: the pendulum starts with angle $\theta = -\frac{\pi}{2}$ without velocity. The models have to predict the dynamics, while no force is applied.

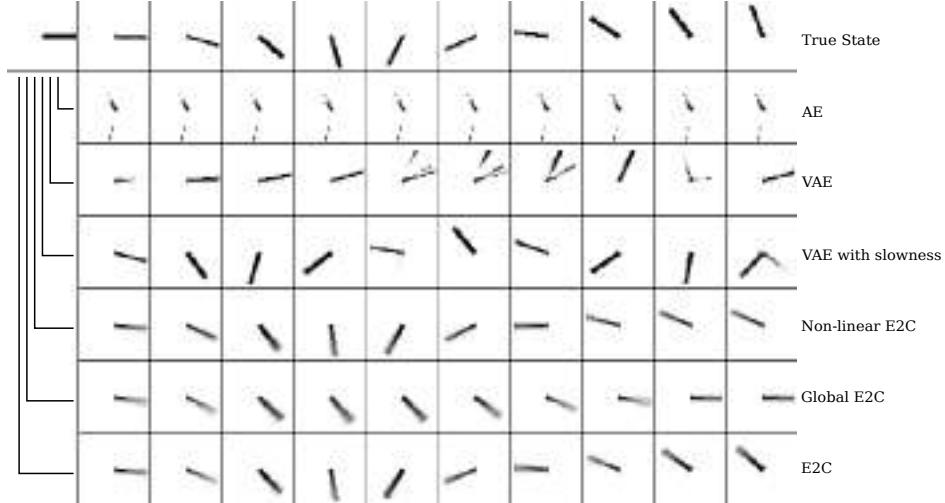


Figure 7: Dreamed trajectories (from left to right) for *controlled* dynamics: the pendulum starts with angle $\theta = \frac{\pi}{2}$ without velocity. For 6 timesteps, full force is applied to the right, followed by 4 timesteps of full force to the left.

C.2 Inverted pendulum latent space

Encoding the pendulum depictions into a 3-dimensional latent space allows for a visual comparison in Figure 8 .

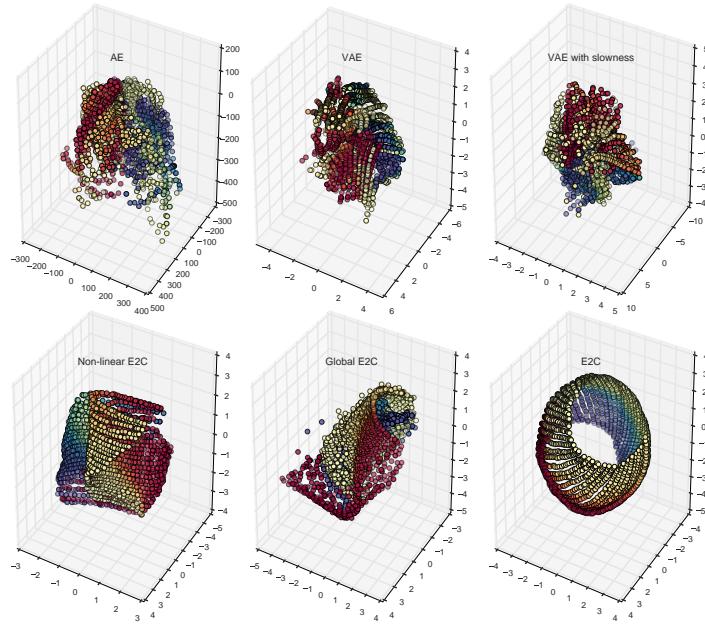


Figure 8: Latent spaces of all baseline models and E2C variants for the inverted pendulum.

C.3 Trajectories for cart-pole and three-link arm

Finally – similar to the images in Section C.1 – Figure 9 shows multi-step predictions for the cart-pole system. We depict important cases: (1) a long-term prediction with the cart-pole standing still (essentially the unstable fix-point of the underlying dynamics); (2) the cart-pole moving to the right, changing the direction of the poles angular velocity (middle column); (3) and the pole moving farthest to the right. The long-term predictions by the E2C model are all of high quality. Note that for the uncontrolled dynamics the predictions show a slight bias of the pole moving to the right (an effect that we consistently saw in trained models for the cart-pole). We attribute this problem to the fact that discretization errors in the image rendering process of the pole angle make it hard to predict small velocities accurately.

C.4 Exemplary trajectory taken for three-link arm task

Figure 10 shows a segment of a controlled trajectory for the three-link arm as executed by the E2C system. Note that, in contrast to other figures in this supplementary material, it does **not** show a long-term prediction but rather 10 steps of a trajectory (together with one-step-ahead predictions) that was taken by the E2C system when combined with model predictive control. For additional visualizations and controlled trajectories for all tasks we refer to the supplementary video.

C.5 Comparison of different models for cart-pole and robot arm

In Table 2 we compare our variety of models in terms of real trajectory cost and task success percentage on the cart-pole and the robot arm. All results are averaged over 30 different starting states with a fixed goal state.

The cart-pole always starts in the goal state (zero angle and zero velocity) with small additive Gaussian noise ($\sigma = 0.01$). Success is defined as preventing the pole from falling below an angle of ± 0.85 rad. The three-link arm system begins in a random configuration and the goal is to to unroll all joints (e.g. make all angles zero) and stay ϵ -close to that position.

The results show that only E2C and its non-linear variant can perform this task successfully, although there is still a large performance gap between the two. We conclude, that the error of linearizing non-linear dynamics after training the corresponding model grows to the point of no longer allowing accurate control for the system.

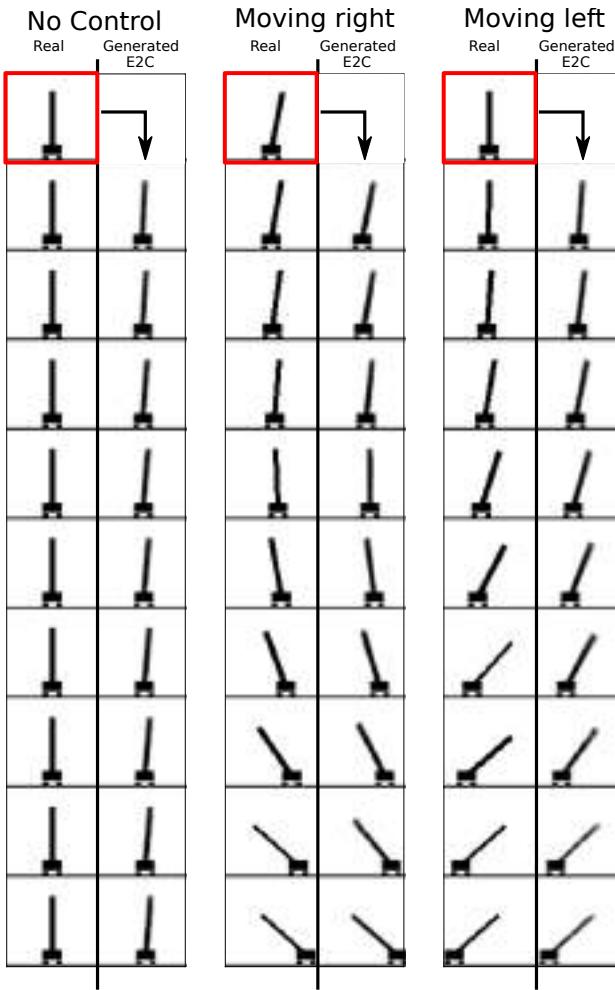


Figure 9: Dreamed trajectories (top to bottom) for uncontrolled (left column) and controlled (middle/right column) dynamics in the cart-pole system. The red image shows the initial configuration, which is encoded resulting in \mathbf{z}_1 . The images in the right half of each column are then generated without additional input by following the dynamics in latent space. The left column depicts the uncontrolled case ($\mathbf{u} = 0$ for all steps). The middle column shows a controlled trajectory with torque -20 applied in each step and the right column a trajectory with torque 20 applied in each step. Prediction of the history image is omitted in these depictions.

Table 2: Comparison between trajectory costs of different approaches for the cart-pole and three-link task. The standard Autoencoder, Variational Autoencoder and Global E2C model are omitted from the table as they failed on this task (performance similar to VAE with slowness).

Algorithm	True model	VAE + slowness	E2C no latent KL	Non-linear E2C	E2C
Cart-Pole balance					
Traj. Cost	15.33 ± 7.70	49.12 ± 16.94	48.90 ± 17.88	31.96 ± 13.26	22.23 ± 14.89
Success %	100 %	0 %	0 %	63 %	93 %
Three-link arm					
Traj. Cost	59.46	1275.53 ± 864.66	1246.69 ± 262.6	460.40 ± 82.18	90.23 ± 47.38
Success %	100 %	0 %	0 %	40 %	90 %

Table 3: Comparison between AICO and iLQR based on the “real” cost for controlling the cart-pole and three-link robot arm using convolutional networks.

Method	iLQR	AICO
Cart-Pole		
E2C	14.56 ± 4.12	12.56 ± 2.47
True model	7.45 ± 1.22	7.03 ± 1.07
Three-Link Robot Arm		
E2C	93.78 ± 32.98	92.99 ± 20.12
True model	53.59 ± 9.74	56.34 ± 10.82

C.6 Comparison of trajectory optimizers for cart-pole and robot arm

To compare how well AICO deals with the covariance matrices estimated in latent space we performed an additional experiment on the cart-pole and three-link robot arm task comparing it to iLQR. We performed model predictive control using the locally linear E2C model starting in 10 different start states each. The remaining settings are as given in Section C.5.

As reported in Table 3, both methods performed about the same for these tasks, indicating that the covariance matrices estimated by our model do not “hurt” planning, but considering them does not improve performance either.

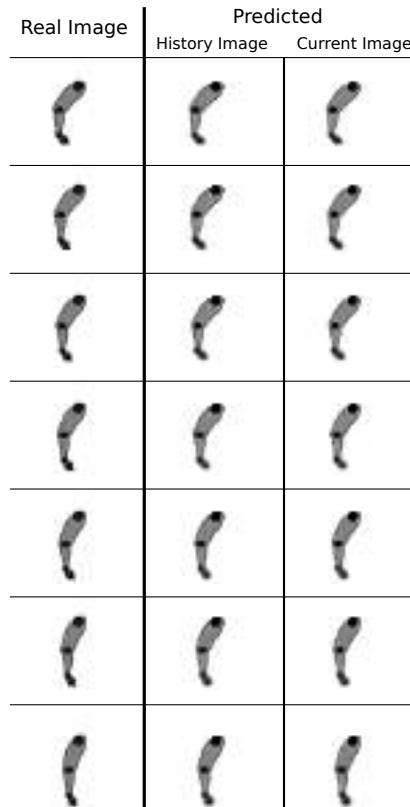


Figure 10: Frames extracted from a trajectory (top to bottom) as executed by the Embed to Control system. The left column shows the real images corresponding to transitions taken in the MDP. Middle and right column show the prediction of history image and current image based on the previous two images.