Shared dynamic model aligned hypernetworks for contextual reinforcement learning

Jan Benad[1], Frank Röder[1], Martin V. Butz[2], Manfred Eppe[1]
Institute for Data Science Foundations, Hamburg University of Technology, Germany[1],
Neuro-Cognitive Modeling Group, University of Tübingen, Germany[2]
jan.benad@tuhh.de

Abstract

We face the challenge of zero-shot generalization in contextual reinforcement learning problems. A distinction is generally made between two cases: either explicit context information is available for the agent, or it is not and has to be inferred from data. We propose DMA*-SH, an approach that builds on dynamic model aligned context inference. It emergently forms context representations and is never informed explicitly about the actual contextual situation it is in. We first show that normalization and random masking can significantly improve the encoded context representation. Second, we enhance context utilization using a hypernetwork which predicts context-dependent weights that are shared between dynamic model, policy, and value function estimation neural modules. Across a diverse set of contextualized environments, we show that our approach achieves superior results, even compared to context-aware baselines.

1 Introduction

Reinforcement Learning (RL) has shown remarkable success in solving complex tasks such as robotic manipulation [Nair et al., 2018] and locomotion [Duan et al., 2016a]. However, RL agents often lack robustness when confronted with variations in task dynamics, such as changes in the mass of objects or surface friction [Moos et al., 2022]. These variations typically require extensive retraining, undermining the generalization capabilities of learned policies [Beck et al., 2023]. This challenge is particularly evident in sim-to-real transfer, where discrepancies between simulation and real-world dynamics can lead to instability and poor performance.

To address this, we propose a method for zero-shot generalization [Kirk et al., 2023] and robust representation learning using Contextual Markov Decision Processes [Hallak et al., 2015, Modi et al., 2018]. In this setup, each *context* corresponds to a distinct variation in transition dynamics, such as altered physical properties (e.g., mass of objects or surface friction). Typically, contextual RL distinguishes two main assumptions: either 1) explicit context information is available as privileged information, or 2) it is not available to the agent, hence it is context-unaware. This work focuses on the latter: we aim to infer the underlying context directly from data, allowing for robust behavior across diverse environments. We extend prior work that encodes a context representation in alignment with a jointly trained dynamic model [Evans et al., 2022, Lee et al., 2020]. We refer to that vanilla baseline as dynamic model aligned (DMA) context inference.

Our contributions. The contributions of this work can be summarized as follows:

- Building on top of recent works for dynamic model aligned context inference [Lee et al., 2020, Evans et al., 2022] for model free contextual RL, we introduce an advanced context encoder architecture DMA* for improved latent context representation achieving superior performance with respect to zero-shot generalization.

18th European Workshop on Reinforcement Learning (EWRL 2025).

- We introduce a novel approach to incorporate dynamic models aligned context information into the agent, using a hypernetwork [Ha et al., 2017] that is trained jointly with the dynamics model and shared with the policy and Q-function. We refer to that as DMA*-SH.
- We introduce a range of contextualized environments making a clear distinction between overlapping and non-overlapping contexts [Beukman et al., 2023]. Especially, the latter are usually unsolvable for simple domain randomization approaches [Tobin et al., 2017], highlighting the necessity for a dedicated context encoder.
- We compare the zero-shot generalization capabilities of our approach to recent methods for contextual RL and obtain superior results. The baselines are comprised of both, *context-aware* – explicit context information is provided as privileged information, and, *context-unaware* agents – explicit context information is not available but is possibly inferred implicitly from past transitions; our case. Aggregated performances are reported with empirical confidence intervals as suggested by Agarwal et al. [2021].

2 Background

Contextual reinforcement learning. We consider a reinforcement learning problem being modeled as a *Markov Decision Process* (MDP). An MDP is defined by a tuple $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, respectively. $P(s'|s, a)$ is the probability of transitioning into state s' after starting in state s and taking action a. $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function and $\gamma \in (0, 1)$ is the discount factor, representing the difference in importance between future and present rewards.

Further, we consider the *Contextual Markov Decision Process* (CMDP) formalism, which is defined by a tuple $(\mathcal{C}, \mathcal{S}, \mathcal{A}, m)$, where $\mathcal{C}$ is the context space, and m is a function that maps a context $c \in \mathcal{C}$ to an MDP $m(c) = (\mathcal{S}, \mathcal{A}, P^c, r^c, \gamma)$. A CMDP thus defines a family of MDPs, that all share an action and state space, but the transition probability P^c and/or the reward function r^c differ depending on the context c. The context c is assumed to be time invariant, i.e., it does not change with time within an episode in the environment. Similar to related work [Beukman et al., 2023, Benjamins et al., 2023, Prasanna et al., 2024] we focus on changes in the transition dynamics P^c and keep the reward function fixed, $r^c = r, \forall c \in \mathcal{C}$.

Zero-shot generalization. Typically, contextual RL is evaluated with respect to zero-shot generalization. Therefore we define three context sets, $\mathcal{C}_{train}$ for training and $\mathcal{C}_{eval,in}$ and $\mathcal{C}_{eval,out}$ for evaluation [Kirk et al., 2023], while $\mathcal{C}_{train} \cap \mathcal{C}_{eval,in} \cap \mathcal{C}_{eval,out} = \emptyset$. Context instances for evaluation are either sampled from the distribution for training contexts $\mathcal{C}_{eval,in}$ or out-of-distribution $\mathcal{C}_{eval,out}$. We are interested in the zero-shot generalization capabilities of the agent, hence, the agent is not allowed to adapt (no gradient updates) to the unknown contexts from $\mathcal{C}_{eval,in}$ and $\mathcal{C}_{eval,out}$.

The agent's objective is to learn a policy π_θ that maximizes the cumulative reward, often expressed as the expected return over the entire training context set $\frac{1}{|\mathcal{C}_{train}|} \sum_c \mathbb{E}_{\pi_\theta} \left[\sum_t \gamma^t r(a_t, s_t) \right]$, where $\mathbb{E}_{\pi_\theta}[\cdot]$ denotes the expectation given that the agent follows policy π_θ and $s_{t+1} \sim P^c(s'|s, a)$ with $c \in \mathcal{C}_{train}$.

3 Related Work

Zero-shot generalization in contextual RL. Contextual RL has been studied in various forms, from cMDPs to domain randomization and meta-RL [Hallak et al., 2015, Modi et al., 2018, Beck et al., 2023]. A recent survey [Kirk et al., 2023] highlights its relevance for zero-shot generalization, emphasizing that separate context sets for training and evaluation enable systematic analysis. One research direction assumes context is observed explicitly as privileged information and integrates it into learning [Chen et al., 2018, Ball et al., 2021, Seyed Ghasemipour et al., 2019, Eghbal-zadeh et al., 2021, Sodhani et al., 2021, Mu et al., 2022, Benjamins et al., 2023, Prasanna et al., 2024]. In contrast, we follow recent work that assumes context can not be observed explicitly. Rather, it is latent and must be inferred [Chen et al., 2018, Xu et al., 2019, Lee et al., 2020, Seo et al., 2020, Xian et al., 2021, Sodhani et al., 2022, Melo, 2022, Evans et al., 2022], focusing on self-supervised context inference through dynamic model alignment. Likely, recurrent agents also create an internal representation of contexts [Grigsby et al., 2024a,b, Luo et al., 2024, Hafner et al., 2019, 2025], although not explicitly being dynamic model aligned.

Related to our work, Beukman et al. [2023] make use of Hypernetworks [Ha et al., 2017] to incorporate context information to the RL models. Still, our approach differs inherently as we do not make the assumption that explicit context information is available.

Meta-RL. Meta-RL trains agents to adapt rapidly to new tasks with minimal experience [Beck et al., 2023], typically by learning adaptive policies that infer task-specific information from past interactions. However, most meta-RL methods require fine-tuning on new tasks [Rakelly et al., 2019, Duan et al., 2016b, Finn et al., 2017, Zintgraf et al., 2019, Nagabandi et al., 2018, Melo, 2022]. Our approach, in contrast, aims for zero-shot generalization by utilizing the latent representations that transfers across environment variations.

Context in Cognition. Besides context approaches in the RL literature, cognitive modeling work has suggested that our minds segment the perceived environment into context-like events [Butz, 2016, Zacks and Tversky, 2001, Zacks et al., 2007, Zacks, 2020]. Along these lines, the recurrent neural network REPRISE was shown to learn latent context representations from scratch, distinguishing between different dynamic regimes [Butz et al., 2019]. More recently, the event-segmentation-oriented perspective has been separated from context. Internally, contextual priors were shown to support the learning of our sensorimotor repertoire as well as other memory structures [Heald et al., 2021, 2023]. Bayesian active inference-based models have shown that context can save computational cognitive effort while modeling human behavior most accurately [Butz, 2022, Cuevas Rivera and Kiebel, 2023, Marković et al., 2021, Mittenbühler et al., 2024, Parr et al., 2023, Schwöbel et al., 2021]. On the deep learning side, contextualized hypernetworks have been introduced in various forms, showing superior generalization and emergent compositionality in early work [Sugita et al., 2011], the emergence of affordance maps [Scholz et al., 2022], as well as the possibility to focus object-oriented encoding pipelines [Traub et al., 2024]. Interlinking neuroscience, developmental psychology, cognitive modeling, and machine learning, a recent interdisciplinary review has pointed out that context inference and context-conditioned learning may be the key to enable behavioral learning in highly complex environments [Butz et al., 2024]—where context invokes task-oriented priors onto both active conceptual model representation and behavioral policies.

4 Context encoding and utilization

In this section we first focus on the representation learning part for a **d**ynamic **m**odel **a**ligned (DMA) context representation and highlight our additions to improve this very representation. We call that improved version DMA*. Then, we describe our novel approach that incorporates latent context information using a shared hypernetwork. We refer to that approach as DMA*-SH, as it extends DMA* with a **s**hared **h**ypernetwork.

4.1 Context inference by dynamic model aligned representation learning

We denote the sliding window of the past K *state-action-next state deltas* transitions $(s_t, a_t, \delta s_{t+1})$, belonging to the same context c as τ_t^c. τ_t^c is fed into the core context encoder $g_\phi(\tau_t^c)$ for which we choose a LSTM layer and its final hidden and cell states are used as context representation z_t. This is in accordance with prior work, where also MLPs, RNNs or Transformer encoder layers were used with slight modifications as the core context encoder [Rakelly et al., 2019, Evans et al., 2022]. Our experiments confirm the experiments performed by Evans et al. [2022] showing that differences in performance are marginal. The gradient updates of the context encoder are driven by the task of learning a representation model. As our contexts solely vary the transition dynamics of the system, a (forward) dynamic model f_θ is sufficient for that task. It predicts the difference between the current and the next state $\delta\hat{s}_{t+1}$ given the current state s_t, action a_t, and the inferred context representation z_t. The model is trained by minimizing a reconstruction loss between the predicted next state difference $\delta\hat{s}_{t+1}$ and the true next state difference δs_{t+1}:

$$L_{\phi,\theta} = \left\| \delta\hat{s}_{t+1} - \delta s_{t+1} \right\|_2 . \tag{1}$$

Given the past transitions τ_t^c we attempt to make the latent z_t as informative as possible for the unknown but underlying contexts c, especially for unseen ones that are out of the training distribution. Prior work [Rakelly et al., 2019, Evans et al., 2022] highlighted that it is beneficial to treat the transitions in τ_t^c in random order, so that the latent states of the context encoder does not contain the temporal structure of τ_t^c. This is an important idea that we adopt. In the following we describe our additional architecture choices for the context encoder. Masking and specific normalization comprise well-known ideas to improve representation learning. To distinguish from the (vanilla) DMA inferred context representation, we refer to our extended approach as DMA* with an emphasis on representation learning.

Input masking. We consider first τ_t^c to be the input to the core context encoder module g_ϕ. Prior works suggest that randomly masking input features or tokens can in general improve representation learning for vision, language and decision making [Devlin et al., 2019, Liu et al., 2022, He et al., 2022]. As we are already relying on an explicit forward dynamics prediction (cf. Equation 1), we do not adopt the prediction task of masked out features which is common in these lines of works. Also, we observe that masking performs best for our purpose with a comparably low masking ratio of 15%. Within τ_t^c we apply random masking on states, actions and next state deltas independently.

Input normalization. After masking, τ_t^c is fed through a linear layer projecting the concatenation of $(s_t, a_t, \delta s_{t+1})$ to a latent model dimension. We continue with a normalization step, for which we experimented with a range of different techniques. Namely, layer normalization [Ba et al., 2016], AvgL1Norm [Fujimoto et al., 2023], SimNorm [Lavoie et al., 2023, Hansen et al., 2024], and a normalization for which statistics are computed across the transitions within τ_t^c (WindowNorm). An ablation is provided in Section A in the Appendix, resulting in best performances choosing AvgL1Norm. Also in theory, AvgL1Norm provides desirable properties: It divides the input vector by its average absolute value in each dimension. With x_i being the i-th dimension of an N-dimensional vector x, then

$$\text{AvgL1Norm}(x) = \frac{x}{\frac{1}{N}\sum_i |x_i|}. \tag{2}$$

AvgL1Norm prevents monotonic growth in the embedding space [Gelada et al., 2019], while keeping the relative scale of the embedding constant during learning without the necessity of updating statistics (as for example in LayerNorm [Ba et al., 2016] or our custom WindowNorm) [Fujimoto et al., 2023].

We tested processing states, actions and next state deltas independently, with no significant benefit. Hence, we omit separate input embeddings for simplicity.

Output normalization. The normalized and masked input τ_t^c is fed into a LSTM layer and the concatenation of its final hidden and cell state are then projected down by a linear layer to a relatively small final dimension for the context representation. We found $z_t \in \mathbb{R}^8$ to be sufficient. Further, we found output normalization to be crucial. Again, we tested different normalization techniques: layer normalization [Ba et al., 2016], AvgL1Norm [Fujimoto et al., 2023] and SimNorm [Lavoie et al., 2023, Hansen et al., 2024] (cf. Section A in the Appendix for an ablation). For DMA*, when the context representation z_t is directly used by the dynamic model and the RL models, best performances were achieved when using SimNorm. Here, the latent representation is normalized by projecting z_t into V-dimensional simplices using a softmax operation. With $z_t \in \mathbb{R}^8$ we are using a smaller $V = 4$. Embedding z_t as simplices promotes sparsity without enforcing discreteness or hard constraints. We refer to Hansen et al. [2024] for further motivation and implementation details.

For DMA*-SH, z_t is used only by an external hypernetwork, hence only indirectly used by the dynamic model and the RL models. In that case we found again AvgL1Norm (cf. Equation 2) to be beneficial. DMA*-SH is described next.

4.2 Context utilization by a shared dynamic model aligned hypernetwork

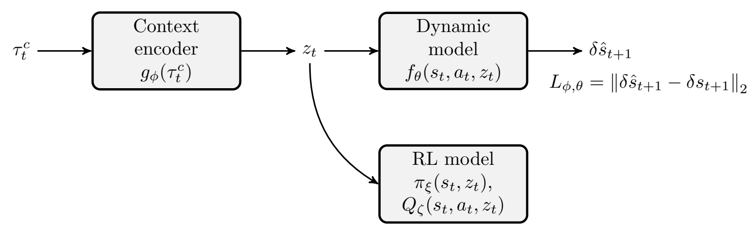
In the (vanilla) DMA case, policy and Q-function simply expect the concatenation of the state s_t with the implicitly inferred context information z_t as input (cf. Figure 1a). In contrast, we are inspired by Beukman et al. [2023] who used a hypernetwork [Ha et al., 2017].

Hypernetworks are meta or second order neural networks [Pollack, 1990, Sugita et al., 2011] that generate weights for a main neural network in an end-to-end differentiable manner.

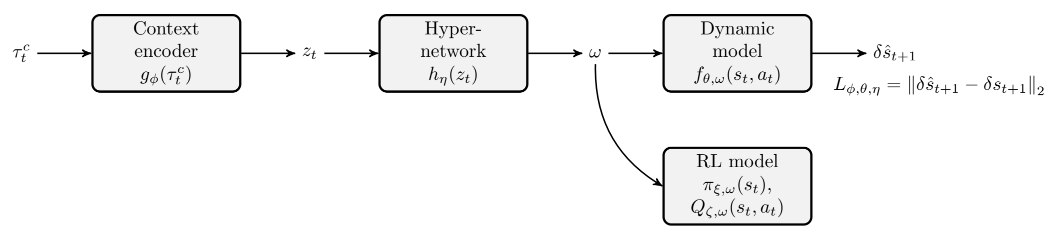
Beukman et al. [2023] assume that the explicit context is available and then condition the hypernetwork on that information to generate weights for parts of the neural networks of the RL models. They call these second order parametrized parts adapters. As we assert that this kind of privileged information often cannot be assumed to be available, our approach takes a detour by inferring first context-like information z_t from past trajectories τ_t^c via the dynamic model aligned representation learning (c.f. Section 4.1). Then, taking z_t as input the parametrized hypernetwork h_η predicts weights ω for a fraction (an adapter) of the neural network for the dynamic model $f_{\theta,\omega}$, hence in total parametrized by ω and the remaining weights θ. The weights ϕ, θ and η for the context encoder, hypernetwork and dynamic model, respectively, are updated jointly with the reconstruction loss

$$L_{\phi,\theta,\eta} = \|\delta\hat{s}_{t+1} - \delta s_{t+1}\|_2 \,. \tag{3}$$

Lastly, without any modifications the generated weights ω are shared with the adapters in the policy $\pi_{\xi,\omega}$ and in the Q-value function $Q_{\zeta,\omega}$. We noticed, performance- and computation-wise this sharing mechanism is more desirable than creating separate hypernetworks for the adapters in the dynamic model and the RL models, being optimized jointly with the respective losses. An overview of the shared hypernetwork approach is provided in Figure 1b. Extending DMA* (cf. Section 4.1) with the shared hypernetwork for context utilization, we refer to our approach as DMA*-SH.



(a) Utilize context representation.



(b) Utilize shared hypernetwork weights.

Figure 1: Schematic overview on how to make use of the inferred context information. (a) Usually the dynamic model aligned context representation z_t is utilized by the RL models [Lee et al., 2020, Evans et al., 2022]. (b) We extend this approach by a hypernetwork h_η whose weights η are updated jointly with the context encoder g_ϕ and the dynamic model $f_{\theta,\omega}$ using the reconstruction loss $L_{\phi,\theta,\eta}$. h_η takes as input the context representation z_t and generates weights ω that are used by the dynamic model and the RL models. When performing the updates for the RL models, gradients through h_η are stopped.

5 Experimental setup

5.1 Metrics

We use a standard evaluation schema for zero-shot generalization in the contextual RL setting [Beukman et al., 2023, Kirk et al., 2023, Benjamins et al., 2023].

We proceed as follows: we sample $n_c = 20$ contexts from the sets $\mathcal{C}_{train}$, $\mathcal{C}_{eval,in}$ and $\mathcal{C}_{eval,out}$, respectively. The agent is trained on the n_c context instances sampled from the training context set $\mathcal{C}_{train}$. Then, for each context we take the trained agent and average its cumulative episodic return over $n_e = 10$ episodes. We then compute the average across contexts within a respective context set. With that we end up with three averaged episodic returns (AER) [Beukman et al., 2023], one for each context set. Performances are reported as AER and as interquartile mean (IQM) with empirical confidence intervals as suggested by Agarwal et al. [2021]. For the latter, we min-max scale them with environment specific upper and lower bounds for the episodic returns, provided in Section 5.3. In general, we run each experiment with $n_s = 10$ different random seed initializations.

5.2 Baselines

For our approaches DMA* and DMA*-SH as well as for all baselines but Amago we use Soft-Actor-Critic [Haarnoja et al., 2018] as the underlying RL algorithm. We do not perform any tuning of SAC's hyperparameters to obtain the best possible comparability. Hyperparameters and implementation details are provided in Section C in the Appendix.

All approaches underlie the same training procedure. The agent is trained in parallel on the $n_c = 20$ contexts drawn from $\mathcal{C}_{train}$.

Domain randomization (DR). This approach has no explicit context information and the agent is not able to infer it. It solely relies on some sort of domain randomization Tobin et al. [2017] as the agent is trained across multiple contexts.

Dynamic model aligned (DMA). Methods such as IIDA [Evans et al., 2022] and CaDM [Lee et al., 2020] rely on dynamic model alignment to represent context information based on recent experience. Usually, the order of transitions used for context inference is random to break temporal correlations (shuffled), and basic dropout is used to improve the latent representation. The latent representation is provided as additional input to policy and Q-function model. Our DMA* extends this line of work, hence we use the vanilla dynamic model alignment as a valid baseline for comparison.

DMA-Pearl. Pearl [Rakelly et al., 2019] is a meta RL algorithm. It uses a probabilistic context encoder to infer context from past transitions. As [Rakelly et al., 2019] tested Pearl solely for reward variations they obtained best results when training the context encoder using gradients from the Bellman updates for the Q-function. Instead we vary the transition dynamics, hence we had to update the context encoder jointly with a dynamic model to achieve comparable baseline performance. We refer to Section A in the Appendix for corresponding ablations. With that, this baseline extends DMA with a probabilistic context encoder and an additional KL loss term to regularize the context representation to a unit Gaussian prior $\mathcal{N}(0, I)$.

Amago. In recurrent agents latent information about the environment can emerge over time (in-context RL). Amago [Grigsby et al., 2024a] is a general purpose in-context RL algorithm for various branches of meta-RL. Although not being solely designed for contextual variations in transition dynamics, it yields a strong baseline using a dedicated recurrent trajectory encoder. For our comparison we use the improved Amago-2 [Grigsby et al., 2024b] with a GRU trajectory encoder.

Concat. This baseline assumes privileged information and concatenates the explicit context with the state. Policy and Q-function expect an expanded state space, $\mathcal{S}' = \mathcal{S} \times \mathcal{C}$. This approach is straight-forward and often the standard approach if explicit context information is available [Ball et al., 2021, Eghbal-zadeh et al., 2021, Sodhani et al., 2021, 2022].

Decision Adapter (DA). Beukman et al. [2023] introduce a strong baseline, again, for the case that context information is explicitly available. But instead of concatenating context with the state, they make use of a hypernetwork architecture inside the policy and optionally the Q-function, where the weights are adapted based on the context. They show strong performance compared to other context-aware baselines such as FLAP [Peng et al., 2021] and cGate [Benjamins et al., 2023].

5.3 Contextualized environments

In the following we describe a range of diverse environments for continuous control that we use to evaluate the agents. To make generalization more difficult we contextualize all environments with two dimensional contexts. A summary of the contextualization with ranges corresponding to the sets $\mathcal{C}_{train}$, $\mathcal{C}_{eval,in}$ and $\mathcal{C}_{eval,out}$ is provided in Table 1. For training we allow 100 000 environment steps per context instance. Note, that we use $n_c = 20$ contexts for training, hence, 2 000 000 environment steps.

We describe the contexts and classify wether they are *i) overlapping* where different context instances are similar enough and an unaware agent without any explicit and implicit knowledge of the context can perform well on average, or *ii) non-overlapping* where such an unaware agent will not be able to solve the task and will perform arbitrarily poorly on average [Beukman et al., 2023].

To obtain true non-overlapping behavior between different context instances, the effect of a varied context has to be drastic w.r.t. the transition dynamics in the environment. For that reason, in some of the listed environments below we allow mirroring the action effect by multiplying the intended action of the agent by a factor of -1, i.e., the action effect is inverted. To illustrate, one might think of the scrolling direction of a computer trackpad or mouse. Depending on the preference, some people prefer congruent behavior, i.e., screen content follows the scrolling direction, and some people prefer the inverted behavior. When being confronted with the non-preferred setting, it is impossible to operate the computer without adaptation (zero-shot) and without being able to infer the dynamics from experience. Contexts are non-overlapping.

DI. We create a custom two dimensional double integrator environment. This version of the environment is frictionless. The agent is represented by a simple mass. It is initialized randomly in the corner positions and its task is to reach the origin at $[0, 0]$ allowing a small margin. The agent is actuated by forces in x, y-direction and its state comprises x, y positions and velocities. The reward signal is sparse, i.e., $+1$ if it reaches the goal position, 0 otherwise. This version of the environment is contextualized by the mass of the agent and by an actuator factor which can either be -1 or 1. The latter context makes it impossible for the agent to solve the task if the agent has neither explicit nor implicit knowledge of the context, i.e., contexts are non-overlapping. Episodic returns are scaled between 0 and 100 (cf. Section 5.1).

DI-friction. Similar to DI, although this version contains friction. It is contextualized by the mass of the agent and by the friction value. Different contextualized environment instances are similar enough and hence overlapping. Episodic returns are scaled between 0 and 100 (cf. Section 5.1).

ODE. Beukman et al. [2023] created this environment to study contextualized RL. It is described by an ordinary differential equation (ODE), parametrized by two context variables c_0 and c_1. The dynamics equation is $x_{t+1} = x_t + \dot{x}_t dt$, with $\dot{x} = c_0 a + c_1 a^2$. For more information, please refer to Beukman et al. [2023]. We observed that an unaware agent performs poorly, hence we argue that context instances are non-overlapping. Episodic returns are scaled between 0 and 200 (cf. Section 5.1).

Cartpole. It is part of the DM control suite (cartpole-balance-v0) [Tassa et al., 2018]. The task is to balance an unactuated pole by applying forces to a cart at its base [Barto et al., 1983]. This environment is contextualized by the pole length and similar to DI by an actuator factor which can either be -1 or 1. Again, hence contexts are non-overlapping. Episodic returns are scaled between 0 and 1000 (cf. Section 5.1).

BallInCup. It is part of the DM control suite (ball_in_cup-catch-v0) [Tassa et al., 2018]. An actuated receptacle can move in the vertical plane in order to swing and catch a ball attached to its bottom. The reward signal is sparse $+1$ if the ball is in the cup, 0 otherwise. The environment is contextualized such that the tendon length and the gravity can be varied. Although, it can be tough to solve for an unaware agent, we consider context instances to rather overlap. Episodic returns are scaled between 0 and 1000 (cf. Section 5.1).

Walker. It is part of the DM control suite (walker-walk-v0) [Tassa et al., 2018]. A planar walker is rewarded for moving forward [Lillicrap et al., 2015]. The contextualization is the same as in the work by Prasanna et al. [2024] where they vary actuator strength (we refer to that strength as an actuator factor) and gravity. It is easily approachable by an unaware agent, hence we consider the contextualized environment instances to overlap. Episodic returns are scaled between 0 and 1000 (cf. Section 5.1).

5.4 Zero-shot generalization

When evaluating our proposed approaches, the main emphasis is on zero-shot generalization capabilities of the agents. As described in Section 2 and 5.1 we distinguish three cases, corresponding to three context sets $\mathcal{C}_{train}$ for training and $\mathcal{C}_{eval,in}$ and $\mathcal{C}_{eval,out}$ for evaluation within- and out-of-distribution. IQM scores aggregated over all considered contextualized environments (cf. 2) suggest that our approaches DMA* and DMA*-SH achieve strong generalization capabilities, especially in the difficult out-of-distribution evaluation case. The main competitor is the context-aware Concat case, which is only surpassed by DMA*-SH in all three context regimes. For the diverse set of environments and types of contextualization DMA*-SH achieves consistently excellent results in terms of AER scores (cf. Table 2). Notably, simple unaware domain randomization is sufficient for the Walker environment, indicating that for some approaches context information (explicit or inferred) can even distract from solving the task. Although not being solely optimized for changes in the transition dynamics, the context-unaware Amago achieves competitive results in most of the

Table 1: Environment contextualization.

Name	Context	Context ranges: Training	Eval-in	Eval-out
DI	mass	$[0.5, 1.5]$	$(0.5, 1.5)$	$[0.1, 0.5) \cup (1.5, 2.0]$
	actuator factor	$\{-1, 1\}$	$\{-1, 1\}$	$\{-1, 1\}$
DI-friction	mass	$[0.5, 1.5]$	$(0.5, 1.5)$	$[0.1, 0.5) \cup (1.5, 2.0]$
	friction	$[0.5, 1.5]$	$(0.5, 1.5)$	$[0.1, 0.5) \cup (1.5, 2.0]$
ODE	c_0	$[-5, 5]$	$(-5, 5)$	$[-10, -5) \cup (5, 10]$
	c_1	$[-5, 5]$	$(-5, 5)$	$[-10, -5) \cup (5, 10]$
Cartpole	length	$[0.3, 0.85]$	$(0.3, 0.85)$	$[0.1, 0.3) \cup (0.85, 2.0]$
	actuator factor	$\{-1, 1\}$	$\{-1, 1\}$	$\{-1, 1\}$
BallInCup	gravity	$[8.0, 12.0]$	$(8.0, 12.0)$	$[1.0, 8.0) \cup (12.0, 20.0]$
	tendon length	$[0.24, 0.36]$	$(0.24, 0.36)$	$[0.1, 0.24) \cup (0.36, 0.5]$
Walker	gravity	$[4.9, 14.7]$	$[4.9, 14.7]$	$[1.0, 4.9) \cup (14.7, 19.6]$
	actuator factor	$[0.5, 1.5]$	$(0.5, 1.5)$	$[0.1, 0.5) \cup (1.5, 2.0]$

environments (cf. Figure 7 in the Appendix for IQM scores whit omitted BallInCup), also in those with non-overlapping contexts, e.g., the DI environment, which cannot be solved by simple domain randomization, as opposed to DI-friction with its overlapping contexts.

DMA-Pearl shows desirable performance compared to the vanilla DMA, indicating a positive impact of the probabilistic context encoder and the KL regularization. Incorporating these design choices into DMA* and DMA*-SH remains for future work.

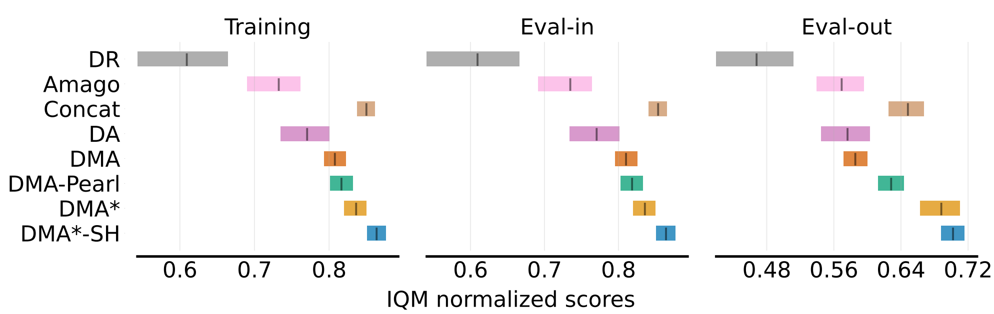


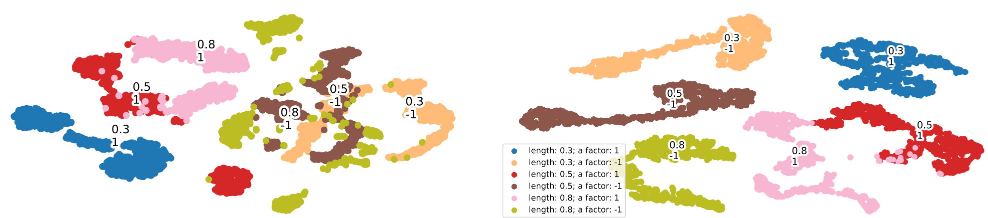
Figure 2: Interquartile mean (IQM) [Agarwal et al., 2021] based on AER scores (cf. Section 5.1) aggregated over the contextualized environments (cf. Section 5.3). We distinguish results for contexts drawn from the three context sets $\mathcal{C}_{train}$, $\mathcal{C}_{eval,in}$ and $\mathcal{C}_{eval,out}$ and compare our approaches DMA* and DMA*-SH to the baselines (cf. Section 5.2).

5.5 Context representation

We evaluate, to what extend our additions in DMA*, namely in- and output normalization and random input masking, improve the context representation z_t compared to the vanilla DMA. Therefore, we contextualize Cartpole (cf. Section 5.3) with a small handcrafted set of contexts. When visualizing z_t using a t-distributed Stochastic Neighbor Embedding (t-SNE) [Van der Maaten and Hinton, 2008], we can observe that DMA* is more capable to distinguish between contexts, which is reflected in more separable clusters in the embedding space (cf. Figure 3). Moreover, when training a simple linear regression model to predict the true contexts based on z_t using the same contextualized example as in Figure 3 we can observe that the context representation from DMA* is more informative than the one from DMA: $R^2 = 92\%$ for DMA* versus $R^2 = 83\%$ for DMA.

Table 2: AER scores (cf. Section 5.1) for each contextualized environment (cf. Section 5.3). Results are averaged across all contexts drawn from the three context sets $\mathcal{C}_{train}$, $\mathcal{C}_{eval,in}$ and $\mathcal{C}_{eval,out}$. We compare our approaches DMA* and DMA*-SH to the baselines (cf. Section 5.2). Best AER scores are highlighted bold. In case multiple approaches are highlighted for an environment, they are within 99% of the maximal achieved AER score. For an simplistic overview we omit variances here. These are reflected in the aggregated IQM visualization (cf. Figure 2). Environment-specific normalization factors are used for the row *Norm. Mean* (cf. Section 5.3).

	Unaware		Aware		Unaware-Inferred			
Name	DR	Amago	Concat	DA	DMA	DMA-Pearl	**DMA***	**DMA*-SH**
DI	22	60	73	38	57	62	71	**76**
DI-friction	61	**79**	70	73	59	65	68	74
ODE	51	168	162	146	157	158	169	**179**
Cartpole	626	619	852	676	**875**	861	**885**	**876**
BallInCup	745	227	862	806	**881**	**885**	**884**	860
Walker	**740**	636	705	708	679	717	651	**733**
Norm. Mean	0.53	0.62	0.78	0.67	0.73	0.75	0.78	**0.81**



(a) DMA.

(b) DMA*.

Figure 3: TSNE visualization [Van der Maaten and Hinton, 2008] comparing the vanilla DMA with the improved DMA*. For visual clarity the Cartpole environment is contextualized with just a few different contexts, listed in the legend and in the center of the corresponding clusters. Pole length and the actuator factor is varied. Each dot corresponds to a z_t encoded from different inputs τ_t^c. For each context we visualize 1000 different encodings. Color coding is based on the true underlying context (unknown for the context encoder).

6 Conclusion

In the domain of contextual RL we consider the assumption that agents are context-unaware and have to infer context information based on past transitions. For the case that the context parametrizes the transition dynamic of the world, this is usually done dynamic model aligned. By applying simple normalization and masking techniques we can improve the context representation significantly (DMA*). Further, we propose a novel approach for utilizing the context representation based on a shared hypernetwork (DMA*-SH). It results in superior zero-shot generalization across a diverse range of contextualized environments, even compared to context-aware methods that assume the explicit context information.

Limitations. Dynamic model aligned methods, hence also our proposed ones, rely on the assumption that solely the transition dynamics are varied. This can be captured by a (forward or inverse) dynamic model. In case the reward function is parametrized by a context (which is allowed in the CMDP formalism), naturally this line of work would not be appropriate. However, a possible solution would be to replace the dynamic model with a reward model. Second, we assume that a context is time-invariant, i.e., it does not change within one episode. We leave considerations to relax that assumption for future works.

References

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G Bellemare. Deep reinforcement learning at the edge of the statistical precipice. *Advances in Neural Information Processing Systems*, 2021.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Philip J Ball, Cong Lu, Jack Parker-Holder, and Stephen Roberts. Augmented world models facilitate zero-shot dynamics generalization from a single offline environment. In *International Conference on Machine Learning*, pages 619–629. PMLR, 2021.

Andrew G Barto, Richard S Sutton, and Charles W Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE transactions on systems, man, and cybernetics*, 1983.

Jacob Beck, Risto Vuorio, Evan Zheran Liu, Zheng Xiong, Luisa Zintgraf, Chelsea Finn, and Shimon Whiteson. A survey of meta-reinforcement learning. *arXiv preprint arXiv:2301.08028*, 2023.

Carolin Benjamins, Theresa Eimer, Frederik Schubert, Aditya Mohan, Sebastian Döhler, André Biedenkapp, Bodo Rosenhahn, Frank Hutter, and Marius Lindauer. Contextualize me – the case for context in reinforcement learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Michael Beukman, Devon Jarvis, Richard Klein, Steven James, and Benjamin Rosman. Dynamics generalisation in reinforcement learning via adaptive context-aware policies. *Advances in Neural Information Processing Systems*, 36:40167–40203, 2023.

Martin V. Butz. Towards a unified sub-symbolic computational theory of cognition. *Frontiers in Psychology*, 7(925), 2016. ISSN 1664-1078. doi: 10.3389/fpsyg.2016.00925.

Martin V. Butz. Resourceful event-predictive inference: The nature of cognitive effort. *Frontiers in Psychology*, 13, 2022. ISSN 1664-1078. doi: 10.3389/fpsyg.2022.867328.

Martin V. Butz, David Bilkey, Dania Humaidan, Alistair Knott, and Sebastian Otte. Learning, planning, and control in a monolithic neural event inference architecture. *Neural Networks*, 117: 135–144, 2019. doi: 10.1016/j.neunet.2019.05.001.

Martin V. Butz, Maximilian Mittenbühler, Sarah Schwöbel, Asya Achimova, Christian Gumbsch, Sebastian Otte, and Stefan Kiebel. Contextualizing predictive minds. *Neuroscience & Biobehavioral Reviews*, 168:105948, 2024. ISSN 0149-7634. doi: https://doi.org/10.1016/j.neubiorev.2024.105948. URL `https://www.sciencedirect.com/science/article/pii/S0149763424004172`.

Tao Chen, Adithyavairavan Murali, and Abhinav Gupta. Hardware conditioned policies for multi-robot transfer learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Dario Cuevas Rivera and Stefan Kiebel. The effects of probabilistic context inference on motor adaptation. *PLOS ONE*, 18(7):1–23, 07 2023. doi: 10.1371/journal.pone.0286749. URL `https://doi.org/10.1371/journal.pone.0286749`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 4171–4186, 2019.

Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. In *International conference on machine learning*, pages 1329–1338. PMLR, 2016a.

Yan Duan, John Schulman, Xi Chen, Peter L Bartlett, Ilya Sutskever, and Pieter Abbeel. Rl^2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016b.

Hamid Eghbal-zadeh, Florian Henkel, and Gerhard Widmer. Context-adaptive reinforcement learning using unsupervised learning of context variables. In *NeurIPS 2020 Workshop on Pre-registration in Machine Learning*, pages 236–254. PMLR, 2021.

Ben Evans, Abitha Thankaraj, and Lerrel Pinto. Context is everything: Implicit identification for dynamics adaptation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2642–2648. IEEE, 2022.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

Scott Fujimoto, Wei-Di Chang, Edward Smith, Shixiang Shane Gu, Doina Precup, and David Meger. For sale: State-action representation learning for deep reinforcement learning. *Advances in neural information processing systems*, 36:61573–61624, 2023.

Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G Bellemare. Deepmdp: Learning continuous latent space models for representation learning. In *International conference on machine learning*, pages 2170–2179. PMLR, 2019.

Jake Grigsby, Linxi Fan, and Yuke Zhu. AMAGO: Scalable in-context reinforcement learning for adaptive agents. In *The Twelfth International Conference on Learning Representations*, 2024a.

Jake Grigsby, Justin Sasek, Samyak Parajuli, Daniel Adebi, Amy Zhang, and Yuke Zhu. AMAGO-2: Breaking the multi-task barrier in meta-reinforcement learning with transformers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b.

David Ha, Andrew M. Dai, and Quoc V. Le. Hypernetworks. In *International Conference on Learning Representations*, 2017.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, 2018.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, 2019.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, pages 1–7, 2025.

Assaf Hallak, Dotan Di Castro, and Shie Mannor. Contextual Markov decision processes. *arXiv preprint arXiv:1502.02259*, 2015.

Nicklas Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations (ICLR)*, 2024.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

James B. Heald, Máté Lengyel, and Daniel M. Wolpert. Contextual inference underlies the learning of sensorimotor repertoires. *Nature*, 600(7889):489–493, 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-04129-3.

James B. Heald, Máté Lengyel, and Daniel M. Wolpert. Contextual inference in learning and memory. *Trends in Cognitive Sciences*, 27(1):43–64, 2023. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2022.10.004. URL https://www.sciencedirect.com/science/article/pii/S1364661322002650.

Christian Henning, Maria R. Cervera, Francesco D'Angelo, Johannes von Oswald, Regina Traber, Benjamin Ehret, Seijin Kobayashi, Benjamin F. Grewe, and João Sacramento. Posterior meta-replay for continual learning. In *Conference on Neural Information Processing Systems*, 2021.

Shengyi Huang, Rousslan Fernand Julien Dossa, Chang Ye, Jeff Braga, Dipam Chakraborty, Kinal Mehta, and João G.M. Araújo. Cleanrl: High-quality single-file implementations of deep reinforcement learning algorithms. *Journal of Machine Learning Research*, 23(274):1–18, 2022. URL `http://jmlr.org/papers/v23/21-1342.html`.

Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel. A survey of zero-shot generalisation in deep reinforcement learning. *Journal of Artificial Intelligence Research*, 76: 201–264, 2023.

Samuel Lavoie, Christos Tsirigotis, Max Schwarzer, Ankit Vani, Michael Noukhovitch, Kenji Kawaguchi, and Aaron Courville. Simplicial embeddings in self-supervised learning and downstream classification. In *International Conference on Learning Representations*, 2023.

Kimin Lee, Younggyo Seo, Seunghyun Lee, Honglak Lee, and Jinwoo Shin. Context-aware dynamics model for generalization in model-based reinforcement learning. In *International Conference on Machine Learning*, pages 5757–5766. PMLR, 2020.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Fangchen Liu, Hao Liu, Aditya Grover, and Pieter Abbeel. Masked autoencoding for scalable and generalizable decision making. *Advances in neural information processing systems*, 35: 12608–12618, 2022.

Fan-Ming Luo, Zuolin Tu, Zefang Huang, and Yang Yu. Efficient recurrent off-policy rl requires a context-encoder-specific learning rate. In *Advances in Neural Information Processing Systems 38*, Vancouver, Canada, 2024.

Dimitrije Marković, Thomas Goschke, and Stefan J. Kiebel. Meta-control of the exploration-exploitation dilemma emerges from probabilistic inference over a hierarchy of time scales. *Cognitive, Affective, & Behavioral Neuroscience*, 21(3):509–533, 2021. ISSN 1531-135X. doi: 10.3758/s13415-020-00837-x. URL `https://doi.org/10.3758/s13415-020-00837-x`.

Luckeciano C Melo. Transformers are meta-reinforcement learners. In *International Conference on Machine Learning*, pages 15340–15359. PMLR, 2022.

Maximilian Mittenbühler, Sarah Schwöbel, David Dignath, Stefan Kiebel, and Martin Butz. A rational trade-off between the costs and benefits of automatic and controlled processing. In *Cognitive Science Conference*. Center for Open Science, 2024. doi: 10.31234/osf.io/gbcxq.

Aditya Modi, Nan Jiang, Satinder Singh, and Ambuj Tewari. Markov decision processes with continuous side information. In *Algorithmic learning theory*, pages 597–618. PMLR, 2018.

Janosch Moos, Kay Hansel, Hany Abdulsamad, Svenja Stark, Debora Clever, and Jan Peters. Robust reinforcement learning: A review of foundations and recent advances. *Machine Learning and Knowledge Extraction*, 4(1):276–315, 2022.

Yao Mu, Yuzheng Zhuang, Fei Ni, Bin Wang, Jianyu Chen, Jianye Hao, and Ping Luo. Domino: Decomposed mutual information optimization for generalized context in meta-reinforcement learning. *Advances in Neural Information Processing Systems*, 35:27563–27575, 2022.

Anusha Nagabandi, Ignasi Clavera, Simin Liu, Ronald S Fearing, Pieter Abbeel, Sergey Levine, and Chelsea Finn. Learning to adapt in dynamic, real-world environments through meta-reinforcement learning. In *International Conference on Learning Representations*, 2018.

Ashvin V Nair, Vitchyr Pong, Murtaza Dalal, Shikhar Bahl, Steven Lin, and Sergey Levine. Visual reinforcement learning with imagined goals. *Advances in neural information processing systems*, 31, 2018.

Thomas Parr, Emma Holmes, Karl J. Friston, and Giovanni Pezzulo. Cognitive effort and active inference. *Neuropsychologia*, 184:108562, 2023. ISSN 0028-3932. doi: 10.1016/j.neuropsychologia.2023.108562. URL `https://www.sciencedirect.com/science/article/pii/S0028393223000969`.

Matt Peng, Banghua Zhu, and Jiantao Jiao. Linear representation meta-reinforcement learning for instant adaptation. *arXiv preprint arXiv:2101.04750*, 2021.

Jordan B Pollack. Recursive distributed representations. *Artificial Intelligence*, 46(1-2):77–105, 1990.

Sai Prasanna, Karim Farid, Raghu Rajan, and André Biedenkapp. Dreaming of many worlds: Learning contextual world models aids zero-shot generalization. *Reinforcement Learning Journal*, 1, 2024.

Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, and Deirdre Quillen. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *International conference on machine learning*, pages 5331–5340. PMLR, 2019.

Fedor Scholz, Christian Gumbsch, Sebastian Otte, and Martin V. Butz. Inference of affordances and active motor control in simulated agents. *Frontiers in Neurorobotics*, 16, 2022. ISSN 1662-5218. doi: 10.3389/fnbot.2022.881673.

Sarah Schwöbel, Dimitrije Marković, Michael N. Smolka, and Stefan J. Kiebel. Balancing control: A bayesian interpretation of habitual and goal-directed behavior. *Journal of Mathematical Psychology*, 100:102472, 2021. ISSN 0022-2496. doi: 10.1016/j.jmp.2020.102472. URL `https://www.sciencedirect.com/science/article/pii/S0022249620301000`.

Younggyo Seo, Kimin Lee, Ignasi Clavera Gilaberte, Thanard Kurutach, Jinwoo Shin, and Pieter Abbeel. Trajectory-wise multiple choice learning for dynamics generalization in reinforcement learning. *Advances in Neural Information Processing Systems*, 33:12968–12979, 2020.

Seyed Kamyar Seyed Ghasemipour, Shixiang Shane Gu, and Richard Zemel. Smile: Scalable meta inverse reinforcement learning through context-conditional policies. *Advances in Neural Information Processing Systems*, 32, 2019.

Shagun Sodhani, Amy Zhang, and Joelle Pineau. Multi-task reinforcement learning with context-based representations. In *International Conference on Machine Learning*, pages 9767–9779. PMLR, 2021.

Shagun Sodhani, Franziska Meier, Joelle Pineau, and Amy Zhang. Block contextual MDPs for continual learning. In *Learning for Dynamics and Control Conference*, pages 608–623. PMLR, 2022.

Yuuya Sugita, Jun Tani, and Martin V Butz. Simultaneously emerging braitenberg codes and compositionality. *Adaptive Behavior*, 19(5):295–316, 2011.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. 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)*, pages 23–30. IEEE, 2017.

Manuel Traub, Frederic Becker, Adrian Sauter, Sebsastian Otte, and Martin V. Butz. Loci-segmented: Improving scene segmentation learning. In Michael Wand, Kristína Malinovská, Jürgen Schmidhuber, and Igor V. Tetko, editors, *Artificial Neural Networks and Machine Learning – ICANN 2024*, pages 45–61, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-72338-4.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Zhou Xian, Shamit Lal, Hsiao-Yu Tung, Emmanouil Antonios Platanios, and Katerina Fragkiadaki. Hyperdynamics: Meta-learning object and agent dynamics with hypernetworks. In *International Conference on Learning Representations*, 2021.

Zhenjia Xu, Jiajun Wu, Andy Zeng, Joshua B Tenenbaum, and Shuran Song. Densephysnet: Learning dense physical object representations via multi-step dynamic interactions. In *Robotics: Science and Systems (RSS)*, 2019.

Jeffrey M. Zacks. Event perception and memory. *Annual Review of Psychology*, 71(1):165–191, 2020. doi: 10.1146/annurev-psych-010419-051101.

Jeffrey M. Zacks and Barbara Tversky. Event structure in perception and conception. *Psychological Bulletin*, 127(1):3–21, 2001. ISSN 1939-1455(Electronic);0033-2909(Print). doi: 10.1037/0033-2909.127.1.3.

Jeffrey M. Zacks, Nicole K. Speer, Khena M. Swallow, Todd S. Braver, and Jeremy R. Reynolds. Event perception: A mind-brain perspective. *Psychological Bulletin*, 133(2):273–293, 2007. doi: 10.1037/0033-2909.133.2.273.

Luisa Zintgraf, Kyriacos Shiarli, Vitaly Kurin, Katja Hofmann, and Shimon Whiteson. Fast context adaptation via meta-learning. In *International conference on machine learning*, pages 7693–7702. PMLR, 2019.

A Ablations

We perform a range of ablations on which we base the design choices in Section 4.1. Figure 4 for DMA* and Figure 5 for DMA*-SH show probability of improvements as suggested by Agarwal et al. [2021]. They only show if there is a likely improvement using our choices compared to the alternatives. They do not necessarily tell us something about the magnitude. In Figure 2 we compare the vanilla DMA to DMA* DMA*-SH, indicating that our design choices cumulatively have significant impact.

In Figure 6 we compare IQM scores Agarwal et al. [2021] for different ratios of the random input masking of actions, states, and next state deltas in τ_t^c resulting in a ratio of 15% to be overall beneficial.

We noticed that the baseline Amago struggles with the BallInCup environment. IQM scores raise significantly when omitting this very environment (cf. Figure 7).

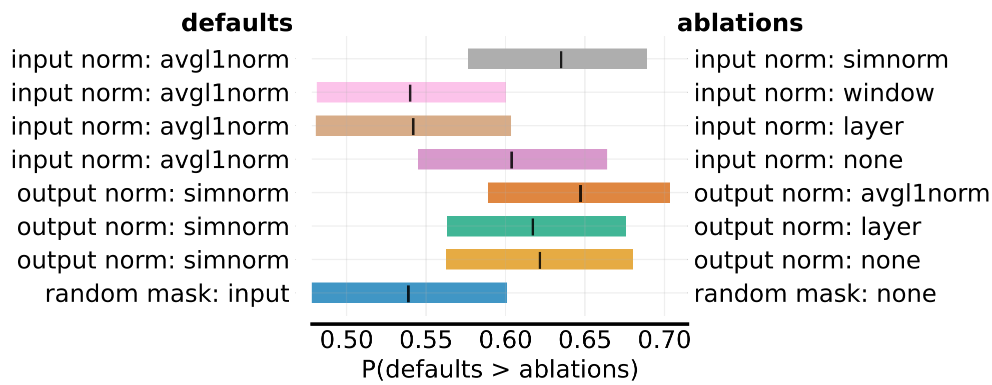


Figure 4: Probability of improvement (POI) [Agarwal et al., 2021] based on AER scores (cf. Section 5.1) aggregated over the contextualized environments (cf. Section 5.3) and over contexts drawn from the three context sets $\mathcal{C}_{train}$, $\mathcal{C}_{eval,in}$ and $\mathcal{C}_{eval,out}$. We ablate the random masking and compare different normalization techniques. POI is based on DMA*, i.e., the usual concatenation of the dynamic model aligned context representation with the state.

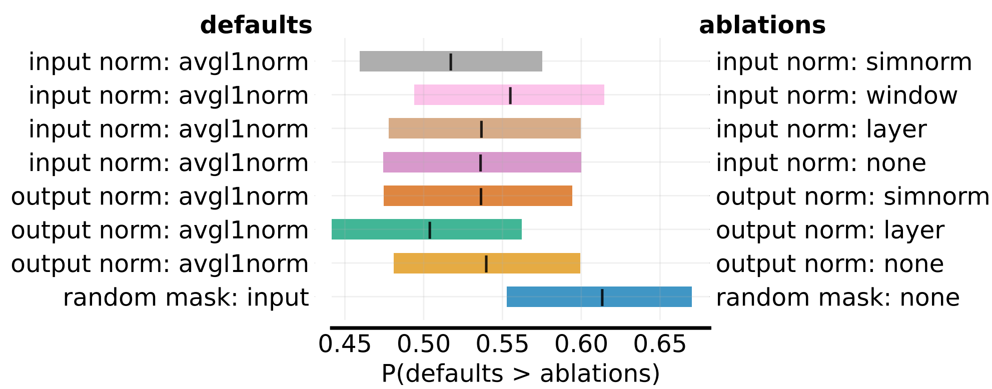


Figure 5: Probability of improvement (POI) [Agarwal et al., 2021] based on AER scores (cf. Section 5.1) aggregated over the contextualized environments (cf. Section 5.3) and over contexts drawn from the three context sets $\mathcal{C}_{train}$, $\mathcal{C}_{eval,in}$ and $\mathcal{C}_{eval,out}$. We ablate the random masking and compare different normalization techniques. POI is based on DMA*-SH, i.e., the novel dynamic model aligned context utilization based on a shared hypernetwork.

B Context representations

In Figure 9 and 10 we provide more examples underlining the results in Section 5.5.

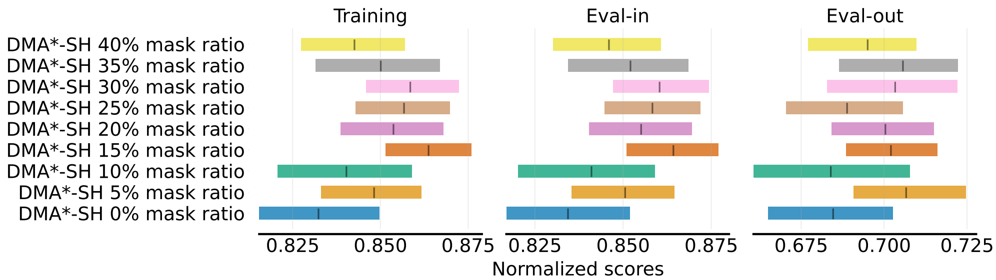


Figure 6: Interquartile mean (IQM) [Agarwal et al., 2021] based on AER scores (cf. Section 5.1) aggregated over the contextualized environments (cf. Section 5.3). We distinguish results for contexts drawn from the three context sets $\mathcal{C}_{train}$, $\mathcal{C}_{eval,in}$ and $\mathcal{C}_{eval,out}$. Using DMA*-SH we compare different ratios for the random input masking. When averaging over the three context sets, best performance is achieved using a ratio of 15%.

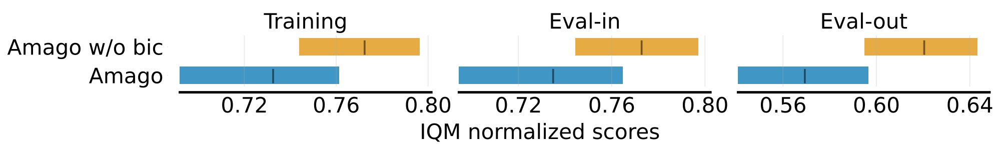


Figure 7: Interquartile mean (IQM) [Agarwal et al., 2021] based on AER scores (cf. Section 5.1) aggregated over the contextualized environments (cf. Section 5.3). We distinguish results for contexts drawn from the three context sets $\mathcal{C}_{train}$, $\mathcal{C}_{eval,in}$ and $\mathcal{C}_{eval,out}$. We notice that Amago struggles with the contextualized BallInCup environment skewing the aggregated performance significantly. As Amago is not explicitly designed for changes in the transition dynamics, we highlight its performance showing aggregated IQM without BallInCup.

C Hyperparameters and implementation details

Table 3 provides an overview for the used hyperparameters of the SAC agent, the context encoder and the dynamic model. We did not perform any tuning for SAC and kept hyperparameters standard as provided in CleanRL [Huang et al., 2022]. We noticed that context window size K of the context encoder depends on the environments. Environments that are originated from the DM Control Suite required a larger K compared to the other ones. The context encoder then takes just a random fraction of the K transitions as input. A relatively small fraction is sufficient. For example in the DM Control Suite case, the context encoder only sees $128 * 0.2 \approx 25$ transitions as input for its τ_t^c.

For our hypernetworks we use the framework by Henning et al. [2021] providing an easy access. The adapter architecture is kept the same as Beukman et al. [2023]. For implementation details we refer to their extended Appendix.

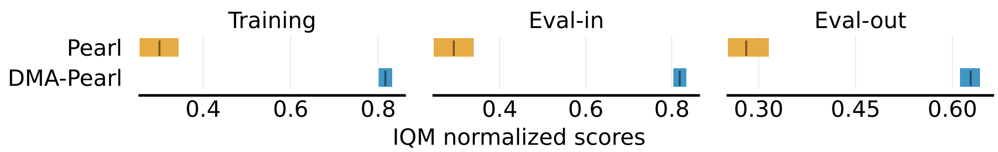
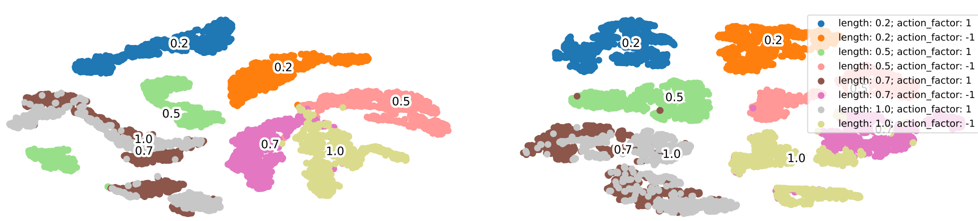


Figure 8: Interquartile mean (IQM) [Agarwal et al., 2021] based on AER scores (cf. Section 5.1) aggregated over the contextualized environments (cf. Section 5.3). We distinguish results for contexts drawn from the three context sets $\mathcal{C}_{train}$, $\mathcal{C}_{eval,in}$ and $\mathcal{C}_{eval,out}$. We compare the original Pearl approach aligned with the Q-function to the dynamic model aligned variant that we are using as a baseline.

(a) DMA.

(b) DMA*.

Figure 9: TSNE visualization [Van der Maaten and Hinton, 2008] comparing the vanilla DMA with the improved DMA*. For visual clarity the Cartpole environment is contextualized with just a few different contexts, listed in the legend and in the center of the corresponding clusters. Pole length and the actuator factor is varied. Each dot corresponds to a z_t encoded from different inputs τ_t^c. For each context we visualize 1000 different encodings. Color coding is based on the true underlying context (unknown for the context encoder). Training a simple linear regression model to predict the true contexts based on z_t we achieve $R^2 = 97\%$ for DMA* and $R^2 = 91\%$ for DMA. Compared to Figure 3 with a different random seed initialization.



(a) DMA.

(b) DMA*.

Figure 10: TSNE visualization [Van der Maaten and Hinton, 2008] comparing the vanilla DMA with the improved DMA*. For visual clarity the Cartpole environment is contextualized with just a few different contexts, listed in the legend and in the center of the corresponding clusters. Pole length and the actuator factor is varied. Each dot corresponds to a z_t encoded from different inputs τ_t^c. For each context we visualize 1000 different encodings. Color coding is based on the true underlying context (unknown for the context encoder). Training a simple linear regression model to predict the true contexts based on z_t we achieve $R^2 = 90\%$ for DMA* and $R^2 = 84\%$ for DMA. Compared to Figure 3 with a different contextualization.

Table 3: Hyperparameters.

Module	Name	Value
SAC	Buffer capacity	1000000
	Batch size	256
	Discount γ	0.99
	Optimizer	Adam
	Critic LR	0.0003
	Actor LR	0.0003
	Temperature LR	0.0003
	Critic soft target update τ	0.005
	Init temperature (SAC)	1.0
	Init temperature (DrQ)	0.1
	Hidden dims	(256, 256)
Context encoder	LR	0.0003
	Model dim	32
	Dropout	0.1
	Context dim	8
	Context window size K (general)	24
	Context window size K (DMC environments)	128
	Context window fraction	0.2
Dynamic model	LR	0.0003
	Hidden dims	(256, 256)