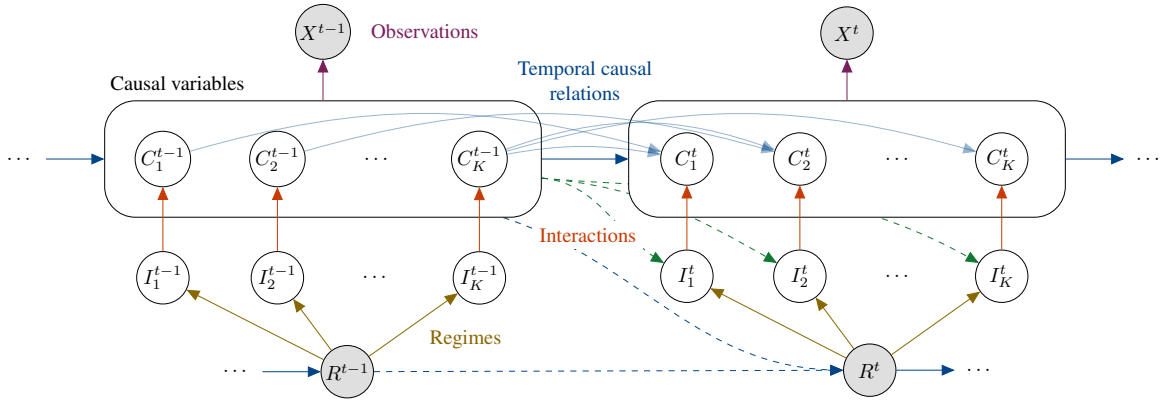

Figure 2: A representation of our assumptions. Observed variables are shown in gray (X^τ and R^τ) and latent variables in white. Optional causal edges are shown as dashed lines. A latent causal variable C_i^t has as parents a subset of the causal factors at the previous time step $C^{t-1} = \{C_1^{t-1}, \dots, C_K^{t-1}\}$, and its latent binary interaction variable I_i^t. The interaction variables are determined by an observed regime variable R^t and potentially by the variables from the previous time step C^{t-1} (*e.g.*, in a collision). The regime variable can be a dynamical process over time as well, for example, by depending on the previous time step. The observation X^τ is a high-dimensional entangled representation of all causal variables C^τ at time step τ.

In this setup, we prove that causal variables are identifiable if the agent interacts with each causal variable in a distinct pattern, *i.e.*, does not always interact with any two causal variables at the same time. We show that for K variables, we can in many cases fulfill this by having as few as $\lfloor \log_2 K \rfloor + 2$ actions with sufficiently diverse effects, allowing identifiability even for a limited number of actions. The binary nature of the interactions permits the identification of a wider class of causal models than previous work in a similar setup, including the common, challenging additive Gaussian noise model (Hyvärinen et al., 1999).

Based on these theoretical results, we propose BISCUIT (**B**inary **I**nteractions for **Cau**sal **I**dentifiability). BISCUIT is a variational autoencoder (Kingma et al., 2014) which learns the causal variables and the agent's binary interactions with them in an unsupervised manner (see Figure 1). In experiments on robotic-inspired datasets, BISCUIT identifies the causal variables and outperforms previous methods. Furthermore, we apply BISCUIT to the realistic 3D embodied AI environment iTHOR (Kolve et al., 2017), and show that BISCUIT is able to generate realistic renderings of unseen causal states in a controlled manner. This highlights the potential of causal representation learning in the challenging task of embodied AI. In summary, our contributions are:

- We show that under mild assumptions, binary interactions with unknown targets identify the causal variables from high-dimensional observations over time.
- We propose BISCUIT, a causal representation learning framework that learns the causal variables and their binary interactions simultaneously.
- We empirically show that BISCUIT identifies both the causal variables and the interaction targets on three robotic-inspired causal representation learning benchmarks, and allows for controllable generations.

2 PRELIMINARIES

In this paper, we consider a causal model $\mathcal{M}$ as visualized in Figure 2. The model $\mathcal{M}$ consists of K latent causal variables $C_1, ..., C_K$ which interact with each other over time, like in a dynamic Bayesian Network (DBN) (Dean et al., 1989; Murphy, 2002). In other words, at each time step t, we instantiate the causal variables as $C^t = \{C_1^t, ..., C_K^t\} \in \mathcal{C}$, where $\mathcal{C} \subseteq \mathbb{R}^K$ is the domain. In terms of the causal graph, each variable C_i^t may be caused by a subset of variables in the previous time step $\{C_1^{t-1}, ..., C_K^{t-1}\}$. For simplicity, we restrict the temporal causal graph to only model dependencies on the previous time step. Yet, as we show in Appendix B.3, our results in this paper can be trivially extended to longer dependencies, *e.g.*, $(C^{t-2}, C^{t-1}) \to C^t$, since C^{t-1} is only used for ensuring conditional independence. As in DBNs, we consider the graph structure to be time-invariant.

Besides the intra-variable dynamics, we assume that the causal system is affected by a regime variable R^t with arbitrary domain $\mathcal{R}$, which can be continuous or discrete of arbitrary dimensionality. This regime variable can model any known external causes on the system, which, for instance, could be a robotic arm interacting with an environment. For the causal graph, we assume that the effect of the regime variable R^t on a causal variable C_i^t can be described by a latent *binary interaction* variable $I_i^t \in \{0, 1\}$. This can be interpreted as each causal variable having two mechanisms/distributions, *e.g.*, an observational and an interventional mechanism, which has similarly been assumed in previous work (Brehmer et al., 2022; Lippe et al., 2022a, 2023). Thereby, the role of the interaction variable I_i^t is to select the mechanism, *i.e.*, observational or interventional, at time step t. For example, a collision between an agent and an object is an interaction that switches the dynamics of the object from