IDENTIFYING TRANSITION-INDEPENDENT STATE VARIABLES FOR FACTORIZED WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning a compact world model representation with factorized state variables is crucial for efficient reinforcement learning. However, existing approaches of identifiable factorization for Markov decision processes (MDPs) are either constrained to block-wise identifiability or rely on the limiting condition of control-independent factors. To extend the capability of identifiable factorization in MDP, this paper proposes Transition-Independent Factorization (TIF), a technique that can further identify factors that have dependent control influence but independent transition dynamics. We first present a theoretical analysis to establish the identifiability of controllable and transition-independent factors, followed by a practical approach to learn factorized world models from high-dimensional observations. Experiments in synthetic worlds and pixel-based grid worlds validate that TIF can accurately identify the ground-truth factors with independent transition dynamics. Experiments in variants of the MiniGrid Empty and Four-Rooms demonstrate that the resulting factorized world model representations can bring substantial benefits to the planning and policy learning in reinforcement learning.

1 INTRODUCTION

The dynamics of complex environments can often be represented by several underlying factors (Higgins et al., 2017b). Imagine a room, shown in Figure 1(a), where the outdoor weather, room brightness, airflow, and sound volume together characterize its current state of interest. Rather than compressing high-dimensional sensing observations into a compact yet entangled latent representation, a factored Markov decision process (MDP) can decompose the representation into disentangled variables with structured dynamics (Guestrin et al., 2003). Such an explicit structure can be exploited to yield benefits for both model learning and planning (Strehl et al., 2007; Chitnis et al., 2020).

As an important component in model-based reinforcement learning (MBRL), world models predict the future by learning environment dynamics (Hafner et al., 2019; 2020), and can further model factored MDPs by factorizing latent states and their transition dynamics. Existing approaches factorize the state according to its role in decision making, such as task relevance (Fu et al., 2021; Zhang et al., 2025) and controllability (Pan et al., 2024; Wang et al., 2022), or according to the compositional structure of environments, such as objects (Kipf et al., 2020) and attributes (Feng & Magliacane, 2023). However, these empirical works do not address the theoretical problem of identifiability, which concerns whether the true factors reflecting the inner workings of the studied phenomena can be statistically identified. Thus, to further recover the true underlying factors from the observations, recent work has begun to study the identifiability of latent representations.

However, existing identifiable factorization methods are either constrained to block-level identifiability or rely on limiting conditions of independently controllable factors. Some methods use actions (Liu et al., 2023) or external semantic information (Wang & Huang, 2025) to identify latent subspaces with specific functional or semantic roles. For example, actions can be used to distinguish controllable and uncontrollable blocks, thereby separating uncontrollable weather from controllable state variables such as brightness, airflow, and volume. However, their identifiability guarantees remain at the block level and cannot recover the underlying factors within each block. More recent effort moves toward factor-level identifiability by targeting independently controllable factors (Rodríguez-Sánchez et al., 2025). However, when the effects of actions on the factors are no

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

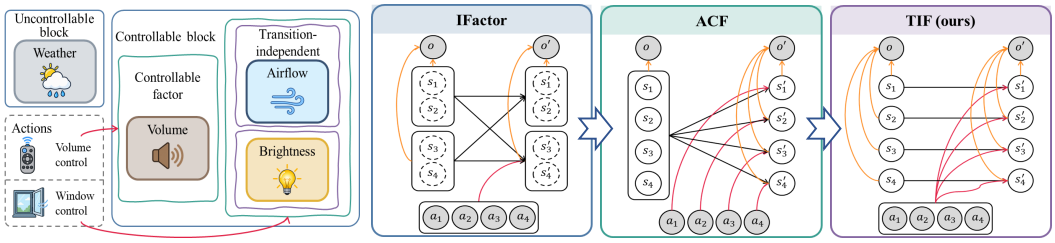


Figure 1: Overview of MDP identifiability. (a) A room environment with uncontrollable and controllable factors, where actions may affect multiple factors. (b) IFactor (Liu et al., 2023): block-level identifiability. (c) ACF (Rodríguez-Sánchez et al., 2025): factor-level identifiability under independently controllable factors. (d) TIF: factor-level identifiability under dependent control influence via factor-wise transition independence.

longer independent—for example, when opening a window simultaneously changes both brightness and airflow—identifiability may degrade to the block level.

We address this identifiability gap by explicitly targeting the recovery of factors that have dependent control influence but independent transition dynamics. Our key idea is to impose transition independency in the world model and learn by predicting interaction trajectories. Thus, even when the same action simultaneously affects multiple factors, sufficiently distinct action-conditioned transitions allow the learned latent factors to recover the corresponding underlying state variables. In our previous room example, this corresponds to recovering brightness and airflow as separate underlying factors, even though the window-control action affects both.

Contributions. First, we establish factor-level identifiability under an additional structural condition—factor-wise transition independence—thereby expanding the scope of MDP identifiability. Second, we propose Transition-Independent Factorization (TIF), a practical method for learning identifiable factors, requiring only a simple replacement of joint latent dynamics in existing world models. Third, we use pixel-based grid worlds to examine the applicability of this condition, and further demonstrate that the learned factorized world models enable more efficient policy learning in two MiniGrid tasks.

2 Problem Formulation

2.1 Transition-Independent Factored MDPs

Consider a class of transition-independent factored Markov Decision Processes (Boutilier et al., 2000) (TI-FMDPs) $\mathcal{M} = \langle S, A, T, R, p_0, \gamma \rangle$ with a state space $S \subseteq \mathbb{R}^K$ and an action space A. The transition function $T : S \times A \to \Delta(S)$ models the world's dynamics, $R : S \times A \to \mathbb{R}$ is a reward function, $\gamma \in [0, 1)$ is the discount factor, and $p_0 \in \Delta(S)$ is the initial state distribution. The agent can interact with the environment to get sequences of observations $\{\langle o_t, a_t, r_t \rangle\}_{t=1}^{T}$. Unlike general MDPs, a TI-FMDP factorizes the state space as $S = S_1 \times \cdots \times S_K$, where $S_i \subseteq \mathbb{R}$, and each factor follows transition-independent dynamics:

$$p(s_t \mid s_{t-1}, a_{t-1}) = \prod_{i=1}^{K} p_i(s_{t,i} \mid s_{t-1,i}, a_{t-1}). \tag{2.1}$$

Importantly, the same action may simultaneously affect multiple factors.

2.2 Nonlinear ICA and Factor Identifiability

We study the identifiability of factors in TI-FMDPs. The term identifiability originates from nonlinear independent component analysis (nonlinear ICA) and refers to the recovery of latent independent generative factors from nonlinearly mixed observations (Hyvärinen et al., 2023; Hyvärinen & Pajunen, 1999). Here, the unobserved latent state $s = (s_1, \ldots, s_K)$ is mixed by an unknown diffeomorphism $f : S \to O$ to generate the observation $o = f(s)$. Given interaction trajectories, we

learn a mapping $g : O \to Z = Z_1 \times \cdots \times Z_K$, where $Z_i \subseteq \mathbb{R}$, that encodes each observation into a factorized latent representation $z = (z_1, \ldots, z_K) = g(o)$. The relationship between the learned representation and the latent state can therefore be written as

$$z = h(s), \qquad h \triangleq g \circ f. \tag{2.2}$$

We say that the learned factorization identifies the underlying factor S_i if there exist a permutation π and invertible functions h_i such that $z_i = h_i(s_{\pi(i)})$ for all $i \in \{1, \ldots, K\}$. Thus, we can recover the underlying factors up to permutation and component-wise invertible transformations.

3 IDENTIFIABILITY THEORY

The core intuition is that actions should induce sufficiently distinct transition responses across different factors. If multiple factors respond similarly to actions, they become difficult to distinguish from interaction data. To formalize this intuition, we first specify a parametric form for the factor transitions and then establish the identifiability result in the following theorem.

We parameterize each factor transition with an M-dimensional conditional exponential family:

$$p_i(s_{t,i} \mid s_{t-1,i}, a_{t-1}) = \frac{b_i(s_{t,i})}{Z_i(s_{t-1,i}, a_{t-1})} \exp\left[\boldsymbol{\phi}_i(s_{t,i})^\top \boldsymbol{\lambda}_i(s_{t-1,i}, a_{t-1})\right]. \tag{3.1}$$

Here, b_i is the base measure, Z_i is the normalizing constant, $\boldsymbol{\phi}_i$ denotes the sufficient statistics, and $\boldsymbol{\lambda}_i$ denotes the corresponding parameters conditioned on its previous state and action. The learned model adopts the same factorized conditional exponential-family form, with $\tilde{\boldsymbol{\phi}}_i$ and $\tilde{\boldsymbol{\lambda}}_i$ denoting its sufficient statistics and corresponding parameters, respectively (Khemakhem et al., 2020).

Theorem 3.1 (Identifiability of Controllable Factors with Transition-Independent Dynamics)**.** *Suppose that both the underlying and learned factors have **independent transition dynamics** defined in Eq. (2.1). If the following conditions hold:*

1. *The observation mapping f and the representation mapping g are C^1 diffeomorphisms.*
2. *For each factor, the sufficient statistics and corresponding parameters are continuously differentiable, and the corresponding conditional and marginal densities are strictly positive in the interior of their supports.*
3. ***Action-response diversity.** For every feasible one-step transition (s_{t-1}, s_t), there exist K distinct actions $a^{(1)}, \ldots, a^{(K)}$ such that the action-response matrix $C(s_{t-1}, s_t) \in \mathbb{R}^{K \times K}$ is full rank, where*

$$C_{ij} = \tilde{\boldsymbol{\phi}}_i'(z_{t,i})^\top \partial_1 \tilde{\boldsymbol{\lambda}}_i \left(z_{t-1,i}, a^{(j)}\right), \qquad z_{t-1} = h(s_{t-1}), \quad z_t = h(s_t). \tag{3.2}$$

4. *The dynamical graph on S, whose edges are the feasible transitions satisfying Condition 3, is connected.*

If the data-generating process and the learned model induce the same interaction trajectory distribution for all admissible action sequences

$$p_O\left(o_{1:T} \mid a_{0:T-1}\right) = \tilde{p}_O\left(o_{1:T} \mid a_{0:T-1}\right), \tag{3.3}$$

then the learned factorization identifies the underlying factors. The proof is provided in Appendix A.

Intuitively, Condition 3 requires different actions to induce sufficiently distinct changes across factors, so that their transitions can be distinguished from interaction trajectories.

Theorem 3.1 reveals two key mechanisms for factor identifiability. Transition independence constrains the structure of latent dynamics, while action-response diversity provides the variation needed to distinguish factor-specific transition mechanisms. Together, they rule out cross-factor mixing in the learned representation. We subsequently use this result to construct a trainable world model.

4 Learning World Models with Transition-Independent Factorization

4.1 World Model Components

Based on the factor identifiability analysis above, we formulate a world model with Transition-Independent Factorization (TIF) to recover identifiable latent factors from interaction trajectories. Each component of the world model is described as follows:

$$\begin{cases} \text{Observation Model:} & p_\theta(o_t \mid z_t), \\ \text{Transition Model:} & p_\gamma(z_t \mid z_{t-1}, a_{t-1}) = \prod_{i=1}^{K} p_{\gamma_i}(z_{t,i} \mid z_{t-1,i}, a_{t-1}), \\ \text{Representation Model:} & q_\phi(z_t \mid o_t, z_{t-1}, a_{t-1}) = \prod_{i=1}^{K} q_{\phi_i}(z_{t,i} \mid o_t, z_{t-1,i}, a_{t-1}). \end{cases} \tag{4.1}$$

The observation model p_θ generates the current observation from the full latent state. The transition model p_γ factorizes the latent dynamics into K transition-independent branches, where each factor depends only on its own previous state and the shared action. This factorization allows a single action to affect multiple latent factors. The representation model q_ϕ serves as an approximate posterior. Each branch receives the full observation but depends on the previous latent state only through the corresponding factor.

Given an initial latent state z_0 and an action sequence $a_{0:T-1}$, the observation and transition models jointly define the trajectory distribution:

$$p_{\theta,\gamma}(o_{1:T}, z_{1:T} \mid z_0, a_{0:T-1}) = \prod_{t=1}^{T} p_\gamma(z_t \mid z_{t-1}, a_{t-1}) p_\theta(o_t \mid z_t). \tag{4.2}$$

Note that transition independency is imposed on the transition model, whereas the representation model adopts a factorized approximate posterior for inference. The latter is an additional modeling assumption and is not implied by the former.

4.2 World Model Estimation

We learn the TIF world model by maximum trajectory likelihood $\log p_{\theta,\gamma}(o_{1:T} \mid z_0, a_{0:T-1})$, consistent with the objective considered in Theorem 3.1. Since this likelihood requires marginalizing over the latent trajectory, we use the representation model to construct the variational posterior

$$Q_\phi(z_{1:T} \mid o_{1:T}, z_0, a_{0:T-1}) = \prod_{t=1}^{T} q_\phi(z_t \mid o_t, z_{t-1}, a_{t-1}). \tag{4.3}$$

and maximize the following evidence lower bound (ELBO): (Kingma & Welling, 2014)

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{Q_\phi}\left[\sum_{t=1}^{T}\left(\log p_\theta(o_t \mid z_t) - \sum_{i=1}^{K} D_{\text{KL}}\Big(q_{\phi_i}(z_{t,i} \mid o_t, z_{t-1,i}, a_{t-1}) \,\|\, p_{\gamma_i}(z_{t,i} \mid z_{t-1,i}, a_{t-1})\Big)\right)\right]. \tag{4.4}$$

The reconstruction term preserves observation information, while the factor-wise KL terms regularize each posterior toward its transition prior. The ELBO provides a tractable approximation to the trajectory likelihood in Theorem 3.1, but optimizing it alone does not guarantee factor identifiability. The full derivation is provided in Appendix B.

Model Instantiation. TIF is independent of the specific policy optimization algorithm used with the learned world model. In this work, we implement TIF based on DreamerV2 (Hafner et al., 2021). As illustrated in Figure 2, we make only a minimal modification: replacing the jointly modeled latent

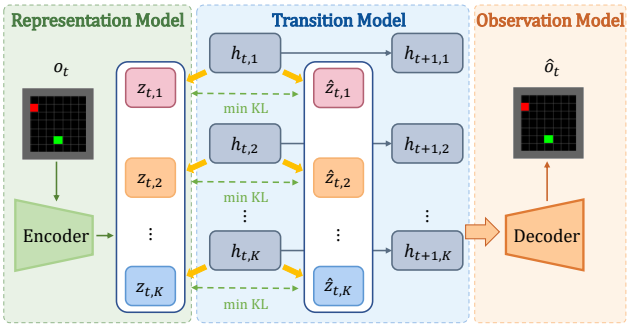


Figure 2: Architecture of the TIF world model with K transition-independent latent factors.

dynamics with transition-independent factorization. As a result, TIF can be readily integrated into existing world models such as IFactor (Liu et al., 2023) and Iso-dream (Pan et al., 2022). According to Theorem 3.1, in addition to transition independency, factor identifiability also relies on sufficiently distinct action responses in the training trajectories. We do not impose action-response diversity as an additional optimization regularizer; rather, it is a property of the interaction data and depends on the coverage of action-conditioned transitions.

5 Evaluation

This section evaluates the factor identifiability of TIF and its benefits for MBRL. We ask three questions: (1) Can TIF recover the controllable and transition-independent factors from observations? (2) What conditions support for factor identifiability? (3) Does identifiable factorization improve the practical utility of world models?

We first evaluate factor recovery in synthetic environments and MiniGrid Empty variants. We then examine the two mechanisms identified by our theory: transition-independent dynamics and action-response diversity. Finally, we evaluate policy learning and long-horizon imagination in variants of MiniGrid Empty and Four-Rooms (Chevalier-Boisvert et al., 2023). Environment descriptions and experimental details are provided in Appendix C and D. The source code is available at here.

5.1 Factor Identifiability

Metrics and Baselines. We evaluate whether TIF can recover the transition-independent factors from interaction trajectories. We use bidirectional R^2 (Bi-R^2) to measure one-to-one correspondence between learned and ground-truth factors, and off-diagonal R^2 to measure cross-factor information leakage. Higher Bi-R^2 and lower off-diagonal R^2 indicate better factor identification. In addition to ACF (Rodríguez-Sánchez et al., 2025), which can identify independently controllable factors, we compare TIF with β-VAE (Higgins et al., 2017a) and FactorVAE (Kim & Mnih, 2018) as unsupervised disentanglement baselines, and with SlowVAE (Klindt et al., 2021), PCL (Hyvärinen & Morioka, 2017), GCL (Hyvärinen et al., 2019), TDRL (Yao et al., 2022), and iVAE (Khemakhem et al., 2020) as temporal or auxiliary-information baselines for identifiable representation learning.

5.1.1 Synthetic Environments

Experiment Setup. We construct a synthetic dynamical system with $K = 5$ continuous scalar factors that satisfies the main identifiability conditions. Each factor follows a transition-independent dynamic conditioned on the action, and $|A| = 8$ discrete actions induce distinct transition responses across factors. The ground-truth state $s_t \in [-5, 5]^5$ is mapped to the observation o_t through an invertible nonlinear mixing function, while actions are sampled independently of the current state.

Results. As shown in Table 1, TIF accurately recovers the ground-truth factors with little cross-factor information leakage, whereas the baselines exhibit either weaker factor recovery or more pronounced factor mixing. These results show that, under controlled settings where the theoretical

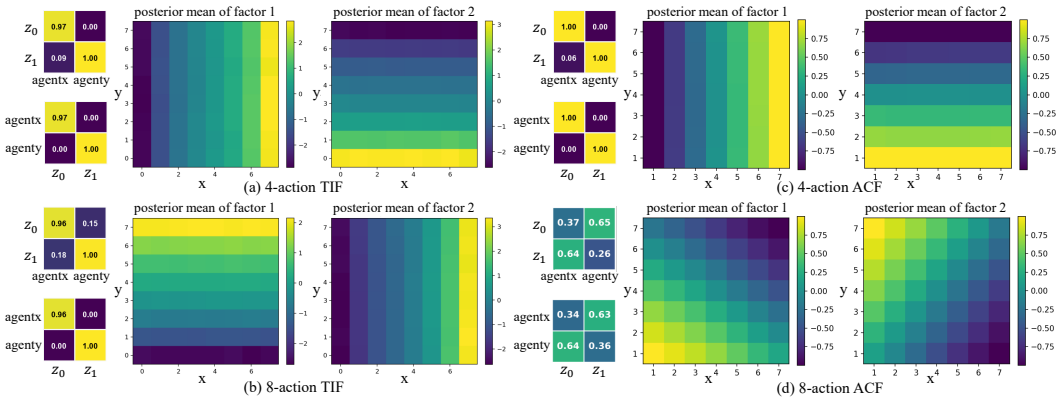


Figure 3: Factor recovery of TIF and ACF in the 4-action and 8-action settings.

conditions hold, TIF can recover the underlying transition-independent factors from nonlinearly mixed observations. Additional identifiability results are provided in Appendix E.

Table 1: Ground-truth factor recovery results in the synthetic environment.

Method	**Bi-R^2 (Mean $\pm$ Std) $\uparrow$**	**Off-diag-R^2 (Mean $\pm$ Std) $\downarrow$**
TIF (ours)	**0.986±0.001**	**0.001±0.001**
ACF	0.707±0.046	0.203±0.051
β-VAE	0.622±0.201	0.326±0.091
FactorVAE	0.665±0.043	0.193±0.023
SlowVAE	0.362±0.022	0.370±0.063
PCL	0.566±0.093	0.078±0.065
GCL	0.792±0.103	0.186±0.066
TDRL	0.684±0.077	0.215±0.041
iVAE	**0.858±0.084**	**0.075±0.063**

Table 2: Ground-truth factor recovery in the Empty variants under different action settings.

Method	**4-action**		**8-action**	
	Bi-R^2 $\uparrow$ (Mean $\pm$ Std)	**Off-diag-R^2 $\downarrow$ (Mean $\pm$ Std)**	**Bi-R^2 $\uparrow$ (Mean $\pm$ Std)**	**Off-diag-R^2 $\downarrow$ (Mean $\pm$ Std)**
TIF (ours)	**0.985±0.018**	**0.023±0.016**	**0.979±0.038**	**0.082±0.089**
ACF	0.916±0.145	0.102±0.121	**0.794±0.183**	0.325±0.114
β-VAE	0.490±0.050	0.316±0.038	0.475±0.082	0.278±0.023
FactorVAE	0.458±0.123	0.221±0.042	0.380±0.053	0.286±0.038
SlowVAE	**0.999±0.000**	**0.014±0.005**	0.602±0.200	0.467±0.035
PCL	0.434±0.066	0.324±0.053	0.628±0.205	0.254±0.097
GCL	0.284±0.056	0.186±0.054	0.327±0.052	0.214±0.061
TDRL	0.492±0.097	0.280±0.110	0.475±0.114	0.298±0.052
iVAE	0.469±0.044	0.278±0.074	0.619±0.119	0.200±0.075

5.1.2 Identifiability from Pixel Inputs

Experiment Setup. To test whether TIF can recover ground-truth factors from visual observations under relaxed theoretical assumptions, we conduct experiments in variations of MiniGrid Empty. The agent moves in a 7×7 grid with ground-truth state $s_t = (x_t^a, y_t^a)$, while the agent observes only $64 \times 64 \times 3$ RGB images and actions. The ground-truth coordinates are used only for evaluation.

Consider two action spaces. The 4-action setting includes up, down, left, and right, while the 8-action setting additionally includes four diagonal actions that can change both ground-truth factors simultaneously. This comparison tests whether factor recovery relies on single-factor changes per action (Rodríguez-Sánchez et al., 2025; Klindt et al., 2021). We further introduce a random-goal variant to examine the identifiability boundary with uncontrollable factors.

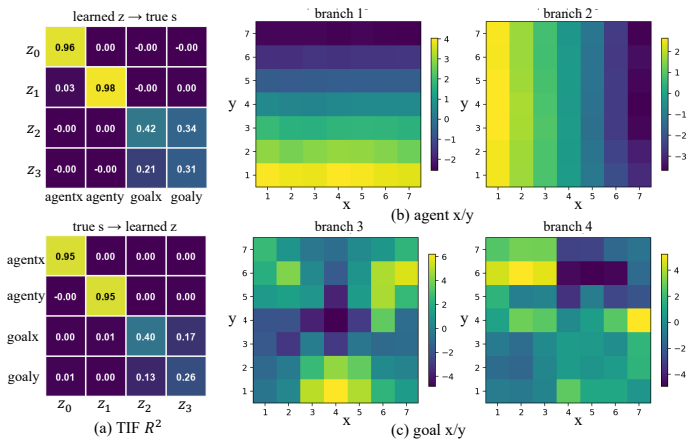


Figure 4: Factor recovery in the random-goal setting with uncontrollable goal factors.

Results. As shown in Table 2, TIF consistently recovers ground-truth position factors in both the 4-action and 8-action settings, with little cross-factor information leakage. Notably, as is shown in Figure 3 and 10, both ACF and SlowVAE recover the agent's two coordinates well in the 4-action setting. With diagonal actions that can simultaneously change both factors, however, their respective assumptions of independent controllability and single-factor transitions are violated, resulting in irresovable entanglement between the learned factors. In contrast, TIF maintains a stable factor correspondence across both action spaces. This suggests that TIF does not rely on the restrictive assumption that each action affects only one ground-truth factor.

Uncontrollable factors. We further consider a random-goal variant with uncontrollable goal factors. The goal coordinates are sampled at the beginning of each episode and remain fixed. As shown in Figure 4, TIF recovers the controllable agent coordinates, while the goal coordinates remain mixed. This illustrates the identifiability boundary of TIF: controllable factors are recoverable under sufficient action-response diversity, whereas uncontrollable factors are not.

5.2 Key Conditions for Factor Identifiability

In this section, we examine two key conditions of TIF: transition-independent dynamics and action-response diversity. The former provides the structural constraint on factor transitions, while the latter supplies the variation needed to distinguish different factors.

5.2.1 Effect of transition-independent dynamics

For transition-independent dynamics, we replace the independent factor-wise transition branches in TIF with a joint transition model. As shown in Table 3, the joint model fails to reliably recover the underlying factors, suggesting that transition independence helps restrict cross-factor mixing.

Table 3: Ablation of transition-independent dynamics in the Empty variant.

Model	**Bi-R^2 (Mean $\pm$ Std) $\uparrow$**	**Off-diag-R^2 (Mean $\pm$ Std) $\downarrow$**
TIF (ours)	**0.985±0.018**	**0.023±0.016**
w/o transition independence	0.743±0.184	0.286±0.119

5.2.2 Effect of Action-Response Diversity

Action-response diversity. We control the rank of the action-response matrix C in the synthetic environment while keeping all other settings fixed. As shown in Figure 5, a full-rank matrix yields a one-to-one correspondence between learned and ground-truth factors. As the matrix becomes rank-deficient, increasing factor mixing emerges in the learned representation. These results are consistent with Theorem 3.1, showing that action-response diversity supports full factor identification.

Action-conditioned transition coverage. In practical reinforcement learning (RL), action-response diversity also depends on how well the behavior policy covers different action-conditioned transi-

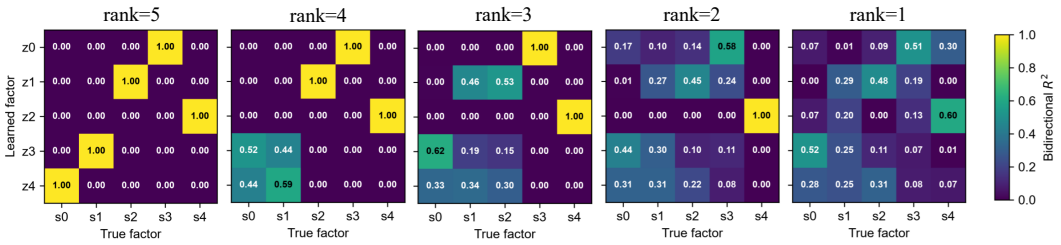


Figure 5: Factor recovery under different ranks of the action-response matrix.

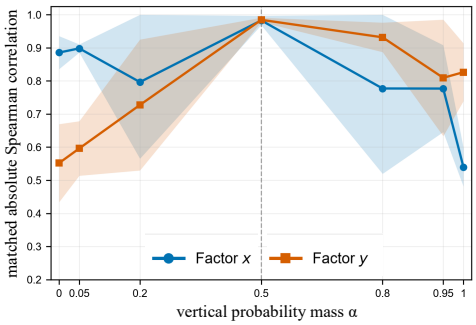


Figure 6: Effect of action-conditioned transition coverage on factor identification.

tions. We vary a parameter α in MiniGrid Empty to control the sampling ratio between horizontal and vertical actions while keeping the environment dynamics and training setup fixed. As shown in Figure 6, both factors are reliably recovered when the two action groups are sampled more evenly. As sampling becomes strongly biased toward one group, recovery of the under-sampled factor degrades markedly. Notably, simply observing all actions is not sufficient; severely imbalanced coverage can still prevent stable identification. This suggests that, in finite-sample RL, action-response diversity is useful only when the relevant action-conditioned transitions are sufficiently covered.

5.3 Benefits of TIF World Models for Policy Learning

Policy learning. We train policies with the learned world models in variants of MiniGrid Empty and Four-Rooms. We compare TIF with DreamerV2 (Hafner et al., 2021) and several factorized world models, including IFactor (Liu et al., 2023), DisWM (Wang et al., 2025), and Iso-Dream (Pan et al., 2022). All methods use comparable model capacity, except for the additionally reported DreamerV2 (large). As shown in Figure 7(a), TIF achieves faster policy learning and stable evaluation returns in both environments. DreamerV2 (large), with increased model capacity, also reaches strong final returns, while TIF retains a clear advantage in policy-learning efficiency.

Imagination. We compare long-horizon imagined trajectories from different world models. Given an initial real observation, each model predicts future states using only the action sequence. TIF maintains more accurate agent dynamics over longer horizons, whereas the other models show larger positional drift. This suggests that identifiable factorization also improves long-horizon imagination.

6 Related Work

Factorized World Models aim to learn structured dynamics by decomposing the latent state and constraining the transition dependencies among latent factors. Existing methods broadly follow function-oriented or environment-oriented factorization. Function-oriented methods separate latent variables according to their roles in decision making, such as reward-relevance (Fu et al., 2021; Huang et al., 2022; Zhang et al., 2025; Huang et al., 2024; Wan et al., 2023) and controllability (Liu et al., 2023; Pan et al., 2024; Wang et al., 2022; Bharadhwaj et al., 2022; Yin et al., 2026), and have been shown to improve sample efficiency in MBRL. Environment-oriented methods instead organize representations according to the compositional or semantic structure of the environment, including object-centric dynamics (Kipf et al., 2020; Feng et al., 2025; Zhang et al., 2026), predefined attributes (Feng & Magliacane, 2023), and external semantic relations (Wang & Huang, 2025).

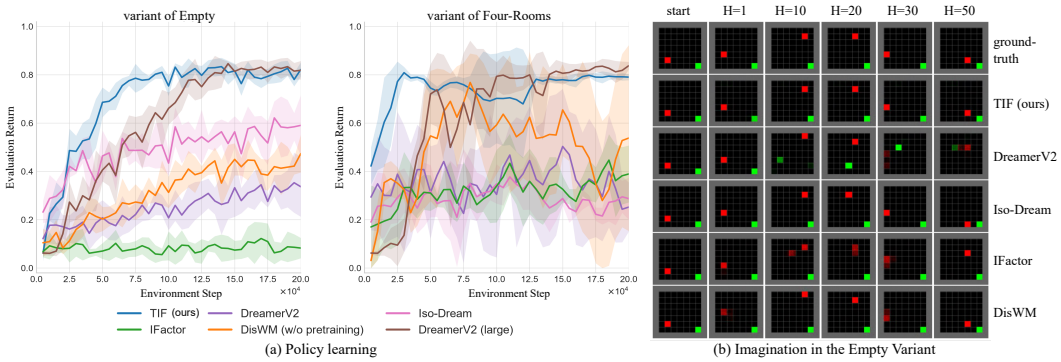


Figure 7: Policy learning and multi-step imagination results in MiniGrid variants.

Auxiliary tasks or inductive biases are introduced to induce these factorizations. Actions (Liu et al., 2023; Pan et al., 2024; Wang et al., 2022; Bharadhwaj et al., 2022) and rewards (Fu et al., 2021; Huang et al., 2022) are used to isolate functional subspaces, while mask-based reconstruction (Huang et al., 2024; Yin et al., 2026; Wan et al., 2023) and dependency regularization (Zhang et al., 2025) encourage separation among latent components. Although effective for learning structured representations, these approaches do not generally guarantee that the learned factors identify the underlying state variables (Locatello et al., 2019), motivating the study of identifiability.

Nonlinear ICA and Identifiability in RL. Nonlinear ICA provides a principled framework for identifying underlying factors. Identifiability typically comes from two aspects: structural constraints on latent generative mechanisms and auxiliary variables for distinguishing factors (Hyvärinen et al., 2023). Within this framework, prior work has established factor-level identifiability under action and state-dependency sparsity (Lachapelle et al., 2022; 2026).

In RL, identifiability studies largely follow the nonlinear ICA perspective, using actions or external semantic information as auxiliary variables while constraining latent dynamics. Early work focuses on block-level identifiability: IFactor (Liu et al., 2023) identifies controllable and reward-relevant blocks using actions and rewards, while WM3C (Wang & Huang, 2025) uses language supervision to recover semantic subspaces. However, factors within each block may remain mixed. More recently, ACF (Rodríguez-Sánchez et al., 2025) achieves factor-level identifiability by imposing sparse action-to-factor effects, enabling recovery of independently controllable factors.

These dynamics-based identifiability results can be viewed through the broader lens of sparsity mechanism (Lachapelle et al., 2026). IFactor exploits sparse relations between auxiliary variables and latent blocks, while ACF extends this perspective to factor-level recovery through action-to-factor effects. In contrast, TIF instead imposes structure on a different dependency relation: factor-to-factor transition independence. It therefore complements existing factor-level identifiability results in MDPs with an additional structural condition based on factor-wise transition independence.

7 Conclusion

This paper proposes Transition-Independent Factorization (TIF) to recover controllable state factors with transition-independent dynamics from the interaction trajectories of MDPs. We prove that even subject to dependent influence of multiple factors from single actions, factor-wise transition independence and sufficiently diverse action responses enable factor-level identification. We establish the corresponding identifiability result and develop a practical method for learning factorized world models. Experiments in synthetic and pixel-based MiniGrid environments show that TIF can accurately recover ground-truth factors and further improve long-horizon prediction and policy learning.

Limitations. Nevertheless, our identifiability result relies on factor-wise independent transition dynamics, whereas real-world factors may exhibit more complex dynamical couplings. Extending TIF to more general dynamical structures and investigating how action-response diversity can benefit policy learning are important directions for future research.

AI USE STATEMENT

In this work, we used generative AI tools to assist with research execution, specifically the implementation of experimental code, and to polish the language of the manuscript. We did not use generative AI tools for research ideation, literature retrieval or discovery, drafting sections of the paper, generating synthetic datasets, or deriving or proving mathematical claims. The mathematical derivations and proofs were carried out by the authors. Other AI-use tasks requiring disclosure were not applicable to this work.

The authors reviewed the AI-assisted text and code. In particular, we checked the polished text for accuracy and fidelity to the intended meaning, and inspected and tested the AI-assisted code and checked the resulting experimental outputs. We take responsibility for the final content of this work, including all text, claims, code, data, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the complete proofs of our theoretical results in the appendix. Detailed descriptions of the synthetic environments and MiniGrid variants, model architectures, training procedures, and hyperparameters are also provided in the appendix. We report results over multiple random seeds and describe the evaluation protocols and metrics used in all experiments. Anonymous source code to reproduce the experiments is provided in the supplementary material.

AUTHOR CONTRIBUTIONS

If you'd like to, you may include a section for author contributions as is done in many journals. This is optional and at the discretion of the authors.

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

Homanga Bharadhwaj, Mohammad Babaeizadeh, Dumitru Erhan, and Sergey Levine. Information prioritization through empowerment in visual model-based RL. In *International Conference on Learning Representations*, 2022.

Craig Boutilier, Richard Dearden, and Moisés Goldszmidt. Stochastic dynamic programming with factored representations. *Artificial Intelligence*, 121(1-2):49–107, 2000.

Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo Perez-Vicente, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan K. Terry. Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. In *Advances in Neural Information Processing Systems*, 2023.

Rohan Chitnis, Tom Silver, Beomjoon Kim, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. Camps: Learning context-specific abstractions for efficient planning in factored mdps. In *Conference on Robot Learning*, volume 155, pp. 64–79, 2020.

Fan Feng and Sara Magliacane. Learning dynamic attribute-factored world models for efficient multi-object reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023.

Fan Feng, Phillip Lippe, and Sara Magliacane. Learning interactive world model for object-centric reinforcement learning. In *Advances in Neural Information Processing Systems*, 2025.

Xiang Fu, Ge Yang, Pulkit Agrawal, and Tommi S. Jaakkola. Learning task informed abstractions. In *International Conference on Machine Learning*, volume 139, pp. 3480–3491, 2021.

Carlos Guestrin, Daphne Koller, Ronald Parr, and Shobha Venkataraman. Efficient solution algorithms for factored mdps. *Journal of Artificial Intelligence Research*, 19:399–468, 2003.

Danijar Hafner, Timothy P. 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*, volume 97, pp. 2555–2565, 2019.

Danijar Hafner, Timothy P. Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

Danijar Hafner, Timothy P. Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. In *International Conference on Learning Representations*, 2021.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017a.

Irina Higgins, Arka Pal, Andrei A. Rusu, Loïc Matthey, Christopher P. Burgess, Alexander Pritzel, Matthew M. Botvinick, Charles Blundell, and Alexander Lerchner. DARLA: improving zero-shot transfer in reinforcement learning. In *International Conference on Machine Learning*, volume 70, pp. 1480–1490, 2017b.

Biwei Huang, Chaochao Lu, Liu Leqi, José Miguel Hernández-Lobato, Clark Glymour, Bernhard Schölkopf, and Kun Zhang. Action-sufficient state representation learning for control with structural constraints. In *International Conference on Machine Learning*, volume 162, pp. 9260–9279, 2022.

Kaichen Huang, Shenghua Wan, Minghao Shao, Hai-Hang Sun, Le Gan, Shuai Feng, and De-Chuan Zhan. Leveraging separated world model for exploration in visually distracted environments. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Aapo Hyvärinen and Hiroshi Morioka. Nonlinear ICA of temporally dependent stationary sources. In *International Conference on Artificial Intelligence and Statistics*, volume 54, pp. 460–469, 2017.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

Aapo Hyvärinen, Hiroaki Sasaki, and Richard E. Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *International Conference on Artificial Intelligence and Statistics*, volume 89, pp. 859–868, 2019.

Aapo Hyvärinen, Ilyes Khemakhem, and Hiroshi Morioka. Nonlinear independent component analysis for principled disentanglement in unsupervised deep learning. *Patterns*, 4(10):100844, 2023.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, volume 108, pp. 2207–2217, 2020.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, volume 80, pp. 2654–2663, 2018.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Thomas N. Kipf, Elise van der Pol, and Max Welling. Contrastive learning of structured world models. In *International Conference on Learning Representations*, 2020.

David A. Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan M. Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. In *International Conference on Learning Representations*, 2021.

Sébastien Lachapelle, Pau Rodríguez, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In *Conference on Causal Learning and Reasoning*, volume 177, pp. 428–484, 2022.

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Nonparametric partial disentanglement via mechanism sparsity: Sparse actions, interventions and sparse temporal dependencies. *Journal of Machine Learning Research*, 27(71):1–90, 2026.

Yuren Liu, Biwei Huang, Zhengmao Zhu, Hong-Long Tian, Mingming Gong, Yang Yu, and Kun Zhang. Learning world models with identifiable factorization. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, volume 97, pp. 4114–4124, 2019.

Minting Pan, Xiangming Zhu, Yunbo Wang, and Xiaokang Yang. Iso-dream: Isolating and leveraging noncontrollable visual dynamics in world models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Minting Pan, Xiangming Zhu, Yitao Zheng, Yunbo Wang, and Xiaokang Yang. Model-based reinforcement learning with isolated imaginations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):2788–2803, 2024.

Rafael Rodríguez-Sánchez, Cameron Allen, and George Dimitri Konidaris. From pixels to factors: Learning independently controllable state variables for reinforcement learning. arXiv preprint arXiv:2510.02484, 2025. URL `https://doi.org/10.48550/arXiv.2510.02484`.

Alexander L. Strehl, Carlos Diuk, and Michael L. Littman. Efficient structure learning in factored-state mdps. In *AAAI Conference on Artificial Intelligence*, pp. 645–650, 2007.

Shenghua Wan, Yucen Wang, Minghao Shao, Ruying Chen, and De-Chuan Zhan. Semail: Eliminating distractors in visual imitation via separated models. In *International Conference on Machine Learning*, volume 202, pp. 35426–35443, 2023.

Qi Wang, Zhipeng Zhang, Baao Xie, Xin Jin, Yunbo Wang, Shiyu Wang, Liaomo Zheng, Xiaokang Yang, and Wenjun Zeng. Disentangled world models: Learning to transfer semantic knowledge from distracting videos for reinforcement learning. In *IEEE/CVF International Conference on Computer Vision*, pp. 2599–2608, 2025.

Tongzhou Wang, Simon S. Du, Antonio Torralba, Phillip Isola, Amy Zhang, and Yuandong Tian. Denoised mdps: Learning world models better than the world itself. In *International Conference on Machine Learning*, volume 162, pp. 22591–22612, 2022.

Xinyue Wang and Biwei Huang. Modeling unseen environments with language-guided composable causal components in reinforcement learning. In *International Conference on Learning Representations*, 2025.

Weiran Yao, Guangyi Chen, and Kun Zhang. Temporally disentangled representation learning. In *Advances in Neural Information Processing Systems*, 2022.

Shicheng Yin, Kaixuan Yin, Weixing Chen, Yang Liu, Guanbin Li, and Liang Lin. DDP-WM: disentangled dynamics prediction for efficient world models. arXiv preprint arXiv:2602.01780, 2026. URL https://doi.org/10.48550/arXiv.2602.01780.

Shaowei Zhang, Jiahan Cao, Dian Cheng, Xunlan Zhou, Shenghua Wan, Le Gan, and De-Chuan Zhan. Leveraging conditional dependence for efficient world model denoising. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Weipu Zhang, Adam Jelley, Trevor McInroe, Amos Storkey, and Gang Wang. Object-centric world models from few-shot annotations for sample-efficient reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 91366–91394, 2026.

A IDENTIFIABILITY OF TIF

Theorem A.1 (Identifiability of Controllable Factors under Transition-Independent Dynamics)**.** *Consider a true state space and a learned state space, each defined as the Cartesian product of K one-dimensional factor spaces,*

$$S = S_1 \times \cdots \times S_K \subseteq \mathbb{R}^K, \qquad Z = Z_1 \times \cdots \times Z_K \subseteq \mathbb{R}^K,$$

where each factor space is a connected open interval.

Suppose that the one-step transition densities of the true and learned models admit factor-wise decompositions:

$$p(s_t \mid s_{t-1}, a_{t-1}) = \prod_{i=1}^{K} p_i(s_{i,t} \mid s_{i,t-1}, a_{t-1}), \tag{A.1}$$

and

$$\widetilde{p}(z_t \mid z_{t-1}, a_{t-1}) = \prod_{i=1}^{K} \widetilde{p}_i(z_{i,t-1} \mid z_{i,t-1}, a_{t-1}). \tag{A.2}$$

Furthermore, assume that the transition density of each factor in both models belongs to an M-dimensional conditional exponential family:

$$p(s_{t,i} \mid s_{t-1,i}, a_{t-1}) = \frac{b_i(s_{t,i})}{Z_i(s_{t-1,i}, a_{t-1})} \exp\left[\boldsymbol{\phi}_i(s_{t,i})^\top \boldsymbol{\lambda}_i(s_{t-1,i}, a_{t-1})\right]. \tag{A.3}$$

and

$$\widetilde{p}(z_{t,i} \mid z_{t-1,i}, a_{t-1}) = \frac{\widetilde{b}_i(z_{t,i})}{\widetilde{Z}_i(z_{t-1,i}, a_{t-1})} \exp\left[\widetilde{\boldsymbol{\phi}}_i(z_{t,i})^\top \widetilde{\boldsymbol{\lambda}}_i(z_{t-1,i}, a_{t-1})\right]. \tag{A.4}$$

Here, $\boldsymbol{\phi}_i$ and $\widetilde{\boldsymbol{\phi}}_i$ denote the vectors of sufficient statistics, whereas $\boldsymbol{\lambda}_i$ and $\widetilde{\boldsymbol{\lambda}}_i$ denote the corresponding parameter vectors determined by the corresponding previous-time state factor and the action. All transition densities are assumed to be strictly positive on the transition supports under consideration.

The true and learned latent states generate the same observation through

$$o = f(s) = \widetilde{f}(z)$$

where $f : S \to O, \quad \widetilde{f} : Z \to O$ are invertible observation maps. This induces the coordinate transformation

$$h \triangleq \widetilde{f}^{-1} \circ f : S \to Z, \qquad z = h(s).$$

If the following conditions hold:

1. *The mapping f and $\widetilde{f}$ are C^1 diffeomorphisms.*
2. *For each factor, the sufficient statistics and corresponding parameters are continuously differentiable, and the corresponding conditional and marginal densities are strictly positive in the interior of their supports.*
3. ***Action-response diversity.** For each factor i, define*

   $$\tilde{\boldsymbol{\phi}}_i'(z_i) \triangleq \frac{d\tilde{\boldsymbol{\phi}}_i(z_i)}{dz_i}, \qquad \partial_1 \tilde{\boldsymbol{\lambda}}_i(z_i, a) \triangleq \frac{\partial \tilde{\boldsymbol{\lambda}}_i(z_i, a)}{\partial z_i}.$$

   For every feasible one-step transition (s_{t-1}, s_t), there exist K distinct actions $a^{(1)}, \ldots, a^{(K)}$ such that the action-response matrix $C(s_{t-1}, s_t) \in \mathbb{R}^{K \times K}$ is full rank, where

$$\boldsymbol{C} = \begin{pmatrix} \widetilde{\boldsymbol{\phi}}_1'(z_{t,1})^\top \partial_1 \widetilde{\boldsymbol{\lambda}}_1(z_{t-1,1}, a^{(1)}) & \cdots & \widetilde{\boldsymbol{\phi}}_K'(z_{t,K})^\top \partial_1 \widetilde{\boldsymbol{\lambda}}_K(z_{t-1,K}, a^{(1)}) \\ \vdots & \ddots & \vdots \\ \widetilde{\boldsymbol{\phi}}_1'(z_{t,1})^\top \partial_1 \widetilde{\boldsymbol{\lambda}}_1(z_{t-1,1}, a^{(K)}) & \cdots & \widetilde{\boldsymbol{\phi}}_K'(z_{t,K})^\top \partial_1 \widetilde{\boldsymbol{\lambda}}_K(z_{t-1,K}, a^{(K)}). \end{pmatrix} \tag{A.5}$$

4. *The dynamical graph on S, whose edges are the feasible transitions satisfying Condition 3, is connected.*

If the data-generating process and the learned model induce the same interaction trajectory distribution for all admissible action sequences

$$p_O\left(o_{1:T} \mid a_{0:T-1}\right) = \tilde{p}_O\left(o_{1:T} \mid a_{0:T-1}\right), \tag{A.6}$$

then the learned factorization identifies the underlying factors up to permutation and componentwise invertible transformations; that is, there exist a permutation π and invertible functions h_i such that

$$z_i = h_i\big(s_{\pi(i)}\big), \qquad \forall i \in \{1, \dots, K\}.$$

Proof. The proof proceeds in four steps:

- **Step 1:** Derive the state-space coordinate transformation and the corresponding one-step transition density relation from observation trajectory equivalence.
- **Step 2:** Substitute the factor-wise conditional exponential-family forms.
- **Step 3:** Take mixed partial derivatives w.r.t. the states in consecutive time steps and use the full-rank condition of the learned action-response matrix to establish a local permutation-diagonal structure.
- **Step 4:** Use the dynamical connectivity to obtain a globally factor-wise mapping.

Step 1: From observational trajectory equivalence to one-step transition equivalence.

Let $\mathbf{a} := a_{1:T-1}$ denote an arbitrary fixed exogenous action sequence. Define the mappings:

$$\begin{aligned} F(s_{1:T}) &:= \big(f(s_1), \dots, f(s_T)\big), \\ \widetilde{F}(z_{1:T}) &:= \big(\widetilde{f}(z_1), \dots, \widetilde{f}(z_T)\big), \\ H(s_{1:T}) &:= \big(h(s_1), \dots, h(s_T)\big). \end{aligned}$$

Since $h = \widetilde{f}^{-1} \circ f$, we have $F = \tilde{F} \circ H$. Therefore, for every true state trajectory $s_{1:T}$,

$$o_{1:T} = F(s_{1:T}) = \widetilde{F}\big(H(s_{1:T})\big). \tag{A.7}$$

Substituting equation A.7 into equation A.6, and applying the change-of-variables formula to the true and learned observation mappings, respectively, gives

$$\begin{aligned} p_o(o_{1:T} \mid \mathbf{a}) &= \widetilde{p}_o(o_{1:T} \mid \mathbf{a}), \\ p_o\big(F(s_{1:T}) \mid \mathbf{a}\big) &= \widetilde{p}_o\big(\widetilde{F}(H(s_{1:T})) \mid \mathbf{a}\big), \\ \frac{p_S(s_{1:T} \mid \mathbf{a})}{|\det J_F(s_{1:T})|} &= \frac{\widetilde{p}_Z(H(s_{1:T}) \mid \mathbf{a})}{\big|\det J_{\widetilde{F}}\big(H(s_{1:T})\big)\big|}. \end{aligned} \tag{A.8}$$

Define

$$q(s_{1:T} \mid \mathbf{a}) := \widetilde{p}_Z(H(s_{1:T}) \mid \mathbf{a}) \prod_{t=1}^{T} |\det J_h(s_t)| \tag{A.9}$$

$$= \widetilde{p}_Z\big(h(s_1), \dots, h(s_T) \mid \mathbf{a}\big) \prod_{t=1}^{T} |\det J_h(s_t)|$$

$$= \widetilde{p}_Z(h(s_1) \mid \mathbf{a}) \, |\det J_h(s_1)| \prod_{t=2}^{T} \left[\widetilde{p}_Z\big(h(s_t) \mid h(s_{t-1}), a_{t-1}\big) \, |\det J_h(s_t)|\right]. \tag{A.10}$$

The last equality follows because the learned model satisfies the Markov property. Hence, q is also a first-order Markov model, with one-step transition density

$$q(s_t \mid s_{t-1}, a_{t-1}) = \widetilde{p}_Z\big(h(s_t) \mid h(s_{t-1}), a_{t-1}\big)\,|\det J_h(s_t)|\,. \tag{A.11}$$

Rearranging equation A.8 and substituting equation A.9 and equation A.10 yields

$$\begin{aligned} p_S(s_{1:T} \mid \mathbf{a}) &= \widetilde{p}_Z(H(s_{1:T}) \mid \mathbf{a}) \frac{|\det J_F(s_{1:T})|}{\big|\det J_{\widetilde{F}}\big(H(s_{1:T})\big)\big|} && \text{(A.12)} \\ &= \widetilde{p}_Z(H(s_{1:T}) \mid \mathbf{a})\,|\det J_H(s_{1:T})| && \text{(A.13)} \\ &= \widetilde{p}_Z(H(s_{1:T}) \mid \mathbf{a}) \prod_{t=1}^{T} |\det J_h(s_t)| && \text{(A.14)} \\ &= q(s_{1:T} \mid \mathbf{a}). && \text{(A.15)} \end{aligned}$$

Here, equation A.13 follows from the chain rule, and equation A.14 uses the time-wise structure of H.

Since the true model p_S and the learned model q have the same joint trajectory density, for any fixed t, marginalizing over the future states $s_{t+1:T}$ gives

$$\begin{aligned} p_S(s_{1:t} \mid \mathbf{a}) &= \int p_S(s_{1:T} \mid \mathbf{a})\, ds_{t+1:T} \\ &= \int q(s_{1:T} \mid \mathbf{a})\, ds_{t+1:T} \\ &= q(s_{1:t} \mid \mathbf{a}). \end{aligned}$$

Further marginalization over s_t gives

$$p_S(s_{1:t-1} \mid \mathbf{a}) = q(s_{1:t-1} \mid \mathbf{a}).$$

Therefore,

$$\begin{aligned} p_S(s_t \mid s_{1:t-1}, \mathbf{a}) &= \frac{p_S(s_{1:t} \mid \mathbf{a})}{p_S(s_{1:t-1} \mid \mathbf{a})} \\ &= \frac{q(s_{1:t} \mid \mathbf{a})}{q(s_{1:t-1} \mid \mathbf{a})} \\ &= q(s_t \mid s_{1:t-1}, \mathbf{a}). \end{aligned} \tag{A.16}$$

Since the action sequence is predetermined, future actions provide no additional information about the current state. Therefore,

$$p_S(s_t \mid s_{1:t-1}, \mathbf{a}) = p_S(s_t \mid s_{1:t-1}, a_{1:t-1}), \tag{A.17}$$

$$q(s_t \mid s_{1:t-1}, \mathbf{a}) = q(s_t \mid s_{1:t-1}, a_{1:t-1}). \tag{A.18}$$

Substituting equation A.17 and equation A.18 into equation A.16, and using the fact that both the true model p_S and the learned model q satisfy the first-order Markov property in equation A.11, we obtain

$$\begin{aligned} p_S(s_t \mid s_{t-1}, a_{t-1}) &= q(s_t \mid s_{t-1}, a_{t-1}) \\ &= \widetilde{p}_Z\big(h(s_t) \mid h(s_{t-1}), a_{t-1}\big)\,|\det J_h(s_t)|\,. \end{aligned} \tag{A.19}$$

Finally, let

$$s' = s_t, \quad s = s_{t-1}, \quad a = a_{t-1}.$$

Suppressing the state-space subscripts when there is no ambiguity, we obtain the following change-of-variables relation for the one-step transition densities:

$$\begin{aligned} p(s' \mid s, a) &= \widetilde{p}\big(h(s') \mid h(s), a\big)\,|\det J_h(s')|\,, \\ \log p(s' \mid s, a) &= \log \widetilde{p}\big(h(s') \mid h(s), a\big) + \log|\det J_h(s')|\,. \end{aligned} \tag{A.20}$$

Equation A.20 transforms observation distributions equally into one-step transition distributions between the true state space and the learned state space.

Step 2: Expanding the Factor-Wise Conditional Exponential-Family Transitions.

Substituting the factorized forms of the one-step transition densities of the true and learned models, i.e. equation A.1 and equation A.2, into equation A.20, gives

$$\sum_{i=1}^{K} \log p_i(s_i' \mid s_i, a) = \sum_{i=1}^{K} \left[\log \widetilde{p}_i\big(h_i(s') \mid h_i(s), a\big)\right] + \log|\det J_h(s')|\,, \tag{A.21}$$

where $h_i(s)$ denotes the i-th output coordinate of the mapping h. At this stage of the proof, we have not assumed that h_i depends only on one true state factor. In general, it may depend on the entire state s. According to the conditional exponential-family assumptions of the true and learned models, substituting equation A.3 and equation A.4 into equation A.21 gives

$$\begin{aligned}
&\sum_{i=1}^{K} \Big[\log b_i(s_i') + \boldsymbol{\phi}_i(s_i')^{\mathsf{T}} \boldsymbol{\lambda}_i(s_i, a) - \log Z_i(s_i, a)\Big] \\
&= \sum_{i=1}^{K} \Big[\log \widetilde{b}_i\big(h_i(s')\big) + \widetilde{\boldsymbol{\phi}}_i\big(h_i(s')\big)^{\mathsf{T}} \widetilde{\boldsymbol{\lambda}}_i\big(h_i(s), a\big) - \log \widetilde{Z}_i\big(h_i(s), a\big)\Big] + \log|\det J_h(s')|\,.
\end{aligned} \tag{A.22}$$

Considering equation A.22, we note the following:

- Terms that depend only on the next state s′:

$$\log b_i(\mathrm{s}_i'), \qquad \log \widetilde{b}_i\big(h_i(\mathrm{s}')\big), \qquad \log|\det J_h(\mathrm{s}')|\,.$$

- Terms that depend only on the current state s and action a:

$$\log Z_i(\mathrm{s}_i, \mathrm{a}), \qquad \log \widetilde{Z}_i\big(h_i(\mathrm{s}), \mathrm{a}\big).$$

- Terms that depend jointly on s′ and s:

$$\boldsymbol{\phi}_i(\mathrm{s}_i')^{\mathsf{T}} \boldsymbol{\lambda}_i(\mathrm{s}_i, \mathrm{a}), \qquad \widetilde{\boldsymbol{\phi}}_i\big(h_i(\mathrm{s}')\big)^{\mathsf{T}} \widetilde{\boldsymbol{\lambda}}_i\big(h_i(\mathrm{s}), \mathrm{a}\big).$$

Step 3: Mixed Partial Derivatives and Local Permutation-Diagonal Structure (Local Identifiability).

We next take mixed partial derivatives of equation A.22 w.r.t. the next-state coordinate s_j' and the current-state coordinate s_k.We use the action-response diversity to show that the Jacobian of h has the same permutation-diagonal structure at both endpoints of every feasible transition.

For the true model, taking the mixed partial derivative gives

$$\frac{\partial^2}{\partial s_j' \partial s_k} \sum_{i=1}^{K} \left[\boldsymbol{\phi}_i(s_i')^{\mathsf{T}} \boldsymbol{\lambda}_i(s_i, a)\right] = \begin{cases} \boldsymbol{\phi}_j'(s_j')^{\mathsf{T}} \partial_1 \boldsymbol{\lambda}_j(s_j, a), & j = k, \\ 0, & j \neq k. \end{cases} \tag{A.23}$$

Here, $\partial_1 \boldsymbol{\lambda}_j(s_j, a)$ denotes the derivative of the corresponding parameter vector w.r.t. its first input s_j. equation A.23 shows that the mixed Hessian of the true transition w.r.t. (s', s) is a diagonal matrix.

For the learned model, only the dynamical coupling term $\widetilde{\boldsymbol{\phi}}_i\big(h_i(s')\big)^{\mathsf{T}} \widetilde{\boldsymbol{\lambda}}_i\big(h_i(s), a\big)$ contributes to the mixed partial derivative. Considering its i-th term, the chain rule gives

$$\begin{aligned}
&\frac{\partial^2}{\partial s_j' \partial s_k} \left[\widetilde{\boldsymbol{\phi}}_i\big(h_i(s')\big)^{\mathsf{T}} \widetilde{\boldsymbol{\lambda}}_i\big(h_i(s), a\big)\right] \\
&\quad = \left[\widetilde{\boldsymbol{\phi}}_i'\big(h_i(s')\big)^{\mathsf{T}} \partial_1 \widetilde{\boldsymbol{\lambda}}_i\big(h_i(s), a\big)\right] \frac{\partial h_i(s')}{\partial s_j'} \frac{\partial h_i(s)}{\partial s_k}.
\end{aligned} \tag{A.24}$$

Define the local response function of the i-th learned factor as

$$c_i(s', s, a) = \boldsymbol{\phi}_i'\big(h_i(s')\big)^{\mathsf{T}} \partial_1 \widetilde{\boldsymbol{\lambda}}_i\big(h_i(s), a\big). \tag{A.25}$$

Comparing equation A.23 and equation A.24, for $\forall j \neq k$, the mixed partial derivative on the true-model side is zero. Therefore, the learned-model side must satisfy

$$\sum_{i=1}^{K} \left[c_i(s', s, a) \frac{\partial h_i(s')}{\partial s_j'} \frac{\partial h_i(s)}{\partial s_k} \right] = 0, \qquad j \neq k. \tag{A.26}$$

Fix an arbitrary feasible transition edge (s, s') and take K actions $a^{(1)}, \ldots, a^{(K)}$ that satisfies Condition 3. For fixed $j \neq k$, define the vector

$$d_{jk}(s', s) = \begin{pmatrix} \dfrac{\partial h_1(s')}{\partial s_j'} \dfrac{\partial h_1(s)}{\partial s_k} \\ \vdots \\ \dfrac{\partial h_K(s')}{\partial s_j'} \dfrac{\partial h_K(s)}{\partial s_k} \end{pmatrix} \in \mathbb{R}^K. \tag{A.27}$$

Define the response matrix

$$C(s', s) = \begin{pmatrix} c_1(s', s, a^{(1)}) & \cdots & c_K(s', s, a^{(1)}) \\ \vdots & \ddots & \vdots \\ c_1(s', s, a^{(K)}) & \cdots & c_K(s', s, a^{(K)}) \end{pmatrix}. \tag{A.28}$$

Applying equation A.26 separately to the actions $a^{(1)}, \ldots, a^{(K)}$, and stacking the K equations, gives

$$C(s', s) d_{jk}(s', s) = 0, \qquad j \neq k. \tag{A.29}$$

By the action-response diversity condition in Condition 3, the matrix $C(s', s)$ is full rank. Therefore, equation A.29 has only the zero solution:

$$d_{jk}(s', s) = 0, \qquad j \neq k.$$

Equivalently, for $\forall i \in \{1, \ldots, K\}$ and $\forall j \neq k$,

$$\frac{\partial h_i(s')}{\partial s_j'} \frac{\partial h_i(s)}{\partial s_k} = 0. \tag{A.30}$$

Since h is a C^1 diffeomorphism, $J_h(s')$ is invertible. Therefore, for every output coordinate i, the i-th row of $J_h(s')$ contains at least one nonzero element. Denote the column containing this element by j_i, i.e.,

$$\frac{\partial h_i(s')}{\partial s_{j_i}'} \neq 0. \tag{A.31}$$

In equation A.30, set $j = j_i$. For $\forall k \neq j_i$, we have

$$\frac{\partial h_i(s')}{\partial s_{j_i}} \frac{\partial h_i(s)}{\partial s_k} = 0.$$

Combining this relation with equation A.31 gives

$$\frac{\partial h_i(s)}{\partial s_k} = 0, \qquad k \neq j_i. \tag{A.32}$$

On the other hand, since $J_h(s)$ is invertible, its i-th row cannot be identically zero. Therefore, the only potentially nonzero element in this row must satisfy

$$\frac{\partial h_i(s)}{\partial s_{j_i}} \neq 0. \tag{A.33}$$

Using equation A.30, now set $k = j_i$. For $\forall j \neq j_i$

$$\frac{\partial h_i(s')}{\partial s_j} \frac{\partial h_i(s)}{\partial s_{j_i}} = 0,$$

Combining this relation with equation A.33 yields

$$\frac{\partial h_i(s')}{\partial s'_j} = 0, \qquad j \neq j_i. \tag{A.34}$$

Therefore, for any feasible transition $s \to s'$, the i-th row of $J_h(s)$ and $J_h(s')$ each contains exactly one nonzero element, and this element is located in the same column j_i.

For each i, define

$$\pi_s(i) := j_i.$$

Because $J_h(s)$ is invertible, two distinct rows cannot place their unique nonzero elements in the same column; otherwise, the corresponding rows would be linearly dependent. Hence, the mapping $i \mapsto \pi_s(i)$ defines a permutation of $\{1, \ldots, K\}$. The same argument applies to s'. Since the unique nonzero element in each row is located in the same column at both endpoints of a feasible transition, we have

$$\pi_s = \pi_{s'}. \tag{A.35}$$

Equivalently, for every feasible transition $s \to s'$ satisfying the action-response diversity condition, there exists a common permutation π_s such that

$$J_h(s) = D_s P_{\pi_s}, \qquad J_h(s') = D_{s'} P_{\pi_s}, \tag{A.36}$$

where D_s and $D_{s'}$ are diagonal matrices with nonzero diagonal entries, and P_{π_s} is the permutation matrix corresponding to the permutation π_s. Thus, equation A.35 establishes the local permutation-diagonal structure of h along every feasible transition edge.

Step 4: From Local Structure to Global Factor-Wise Identifiability (Global Identifiability)

According to the state-space connectivity condition, there exists a fixed anchor point $s^{(0)} \in S$ such that, for any state $x \in S$, there is a path composed of feasible transition edges satisfying Condition 3:

$$s^{(0)} \to s^{(1)} \to \cdots \to s^{(m)} = x. \tag{A.37}$$

By Step 3, for any feasible transition edge $s \to s'$ satisfying the action-response diversity condition, the Jacobians at the two endpoints have the same local permutation. Therefore, applying equation A.35 successively along the feasible path in equation A.37, we obtain

$$\pi_{s^{(0)}} = \pi_{s^{(1)}} = \cdots = \pi_{s^{(m)}}. \tag{A.38}$$

Let $\pi := \pi_{s^{(0)}}$. Since $x \in S$ is arbitrary, it follows that

$$\pi_x = \pi, \qquad \forall x \in S.$$

Therefore, the Jacobian of the mapping h has a globally consistent permutation-diagonal structure over the entire state space:

$$J_h(x) = D(x) P_\pi, \qquad \forall x \in S, \tag{A.39}$$

where $D(x)$ is a diagonal matrix with nonzero diagonal entries, and P_π is the permutation matrix corresponding to π.

Equation A.39 is equivalent to

$$\frac{\partial h_i(s)}{\partial x_j} = 0, \qquad j \neq \pi(i), \tag{A.40}$$

and

$$\frac{\partial h_i(s)}{\partial x_{\pi(i)}} \neq 0. \tag{A.41}$$

Because S is a Cartesian product of connected intervals, the i-th output coordinate h_i remains unchanged along every coordinate direction other than $x_{\pi(i)}$. Hence, there exists a one-dimensional function $\psi_i : S_{\pi(i)} \to Z_i$ such that

$$h_i(s) = \psi_i\big(s_{\pi(i)}\big)\,. \tag{A.42}$$

Moreover, since $\partial h_i(s)/\partial x_{\pi(i)} \neq 0$ and $h : S \to Z$ is a global diffeomorphism, each ψ_i is a one-dimensional C^1 diffeomorphism. Therefore, there exists a global permutation π and a collection of one-dimensional C^1 diffeomorphisms $\{\psi_i\}_{i=1}^K$ such that

$$h_i(s) = \psi_i\big(s_{\pi(i)}\big)\,, \qquad i = 1, \ldots, K. \tag{A.43}$$

This proves that the learned state identifies the true state factors up to a global permutation and component-wise invertible transformations. □

B ELBO FOR TIF

This section derives the variational objective of TIF. Given an observation sequence $o_{1:T}$, an action sequence $a_{0:T-1}$, and a fixed initial latent state z_0, TIF follows the generative model below:

$$p_\theta\big(o_{1:T}, z_{1:T} \mid z_0, a_{0:T-1}\big) = \prod_{t=1}^{T} p_\theta\big(o_t \mid z_t\big) p_\gamma\big(z_t \mid z_{t-1}, a_{t-1}\big), \tag{B.1}$$

where $p_\theta(o_t \mid z_t)$ denotes the observation model, and $p_\theta(z_t \mid z_{t-1}, a_{t-1})$ denotes the latent-state transition model.

To approximate the true posterior distribution, we introduce the following variational posterior:

$$Q_\phi\big(z_{1:T} \mid o_{1:T}, z_0, a_{0:T-1}\big) = \prod_{t=1}^{T} q_\phi\big(z_t \mid o_t, z_{t-1}, a_{t-1}\big). \tag{B.2}$$

Each component of the world model is described as follows:

$$\begin{cases} \text{Observation Model:} & p_\theta(o_t \mid z_t), \\ \text{Transition Model:} & p_\gamma(z_t \mid z_{t-1}, a_{t-1}) = \displaystyle\prod_{i=1}^{K} p_{\gamma_i}(z_{t,i} \mid z_{t-1,i}, a_{t-1}), \\ \text{Representation Model:} & q_\phi(z_t \mid o_t, z_{t-1}, a_{t-1}) = \displaystyle\prod_{i=1}^{K} q_{\phi_i}(z_{t,i} \mid o_t, z_{t-1,i}, a_{t-1}). \end{cases} \tag{B.3}$$

The latent state at time t is written componentwise as

$$z_t = \big(z_{t,1}, \ldots, z_{t,K}\big). \tag{B.4}$$

B.1 ELBO

The goal of the world model is to maximize the log-likelihood of the observation trajectory conditioned on the action sequence and the fixed initial latent state:

$$\log p_{\theta,\gamma}\big(o_{1:T} \mid z_0, a_{0:T-1}\big). \tag{B.5}$$

Integrating over the latent trajectory $z_{1:T}$ and introducing the variational posterior in equation B.2, we have

$$\begin{aligned} &\log p_{\theta,\gamma}\big(o_{1:T} \mid z_0, a_{0:T-1}\big) \\ &\quad = \log \int p_\theta\big(o_{1:T}, z_{1:T} \mid z_0, a_{0:T-1}\big)\, \mathrm{d}z_{1:T} \\ &\quad = \log \int Q_\phi\big(z_{1:T} \mid o_{1:T}, z_0, a_{0:T-1}\big) \frac{p_{\theta,\gamma}\big(o_{1:T}, z_{1:T} \mid z_0, a_{0:T-1}\big)}{Q_\phi\big(z_{1:T} \mid o_{1:T}, z_0, a_{0:T-1}\big)} \mathrm{d}z_{1:T} \\ &\quad = \log \mathbb{E}_{Q_\phi}\left[\frac{p_{\theta,\gamma}\big(o_{1:T}, z_{1:T} \mid z_0, a_{0:T-1}\big)}{Q_\phi\big(z_{1:T} \mid o_{1:T}, z_0, a_{0:T-1}\big)}\right]. \end{aligned} \tag{B.6}$$

Applying Jensen's inequality gives the evidence lower bound:

$$\begin{aligned}&\log p_{\theta,\gamma}\big(o_{1:T} \mid z_0, a_{0:T-1}\big)\\ &\quad\geq \mathbb{E}_{Q_\phi}\Big[\log p_{\theta,\gamma}\big(o_{1:T}, z_{1:T} \mid z_0, a_{0:T-1}\big) - \log Q_\phi\big(z_{1:T} \mid o_{1:T}, z_0, a_{0:T-1}\big)\Big]\\ &\quad:= \mathcal{L}_{\mathrm{ELBO}}.\end{aligned} \tag{B.7}$$

Substituting the generative model in equation B.1 and the variational posterior in equation B.2 into equation B.7, the trajectory-level ELBO becomes

$$\begin{aligned}&\mathcal{L}_{\mathrm{ELBO}}\\ &= \mathbb{E}_{Q_\phi}\Big[\sum_{t=1}^{T}\log p_\theta\big(o_t \mid z_t\big) + \sum_{t=1}^{T}\log p_\gamma\big(z_t \mid z_{t-1}, a_{t-1}\big) - \log q_\phi\big(z_t \mid o_t, z_{t-1}, a_{t-1}\big)\Big]\\ &= \sum_{t=1}^{T}\mathbb{E}_{Q_\phi}\left[\log p_\theta\big(o_t \mid z_t\big)\right] + \sum_{t=1}^{T}\mathbb{E}_{Q_\phi}\left[\log p_\gamma\big(z_t \mid z_{t-1}, a_{t-1}\big) - \log q_\phi\big(z_t \mid o_t, z_{t-1}, a_{t-1}\big)\right].\end{aligned} \tag{B.8}$$

B.2 Single-Step Transition Terms

Let $Q_{\phi,t}(z_t \mid o_{1:t}, z_0, a_{0:t-1})$ denotes the marginal distribution of the latent state at time t induced by the variational posterior.

$$Q_{\phi,t}\big(z_t \mid o_{1:t}, z_0, a_{0:t-1}\big) = \int Q_\phi\big(z_{1:t} \mid o_{1:t}, z_0, a_{0:t-1}\big)\mathrm{d}z_{1:t-1} \tag{B.9}$$

For the observation term in equation B.8, since the observation model depends only on z_t, we have

$$\mathbb{E}_{Q_\phi}\left[\log p_\theta\big(o_t \mid z_t\big)\right] = \mathbb{E}_{Q_{\phi,t}(z_t)}\left[\log p_\theta\big(o_t \mid z_t\big)\right]. \tag{B.10}$$

The single-step transition term in equation B.8 can be further decomposed as

$$\begin{aligned}&\mathbb{E}_{Q_\phi}\Big[\log p_\gamma\big(z_t \mid z_{t-1}, a_{t-1}\big) - \log q_\phi\big(z_t \mid o_t, z_{t-1}, a_{t-1}\big)\Big]\\ &= \mathbb{E}_{Q_{\phi,t-1}(z_{t-1})}\Big[\mathbb{E}_{q_\phi(s_t|o_t,s_{t-1},a_{t-1})}\Big[\log p_\gamma\big(z_t \mid z_{t-1}, a_{t-1}\big) - \log q_\phi\big(z_t \mid o_t, z_{t-1}, a_{t-1}\big)\Big]\Big]\\ &= -\mathbb{E}_{Q_{\phi,t-1}(z_{t-1})}\Big[D_{\mathrm{KL}}\Big(q_\phi\big(z_t \mid o_t, z_{t-1}, a_{t-1}\big)\,\Big\|\,p_\gamma\big(z_t \mid z_{t-1}, a_{t-1}\big)\Big)\Big].\end{aligned} \tag{B.11}$$

Therefore, the ELBO can be written as

$$\begin{aligned}\mathcal{L}_{\mathrm{ELBO}} =& \sum_{t=1}^{T}\mathbb{E}_{Q_{\phi,t}(z_t)}\left[\log p_\theta\big(o_t \mid z_t\big)\right]\\ &- \sum_{t=1}^{T}\mathbb{E}_{Q_{\phi,t-1}(z_{t-1})}\Big[D_{\mathrm{KL}}\Big(q_\phi\big(z_t \mid o_t, z_{t-1}, a_{t-1}\big)\,\Big\|\,p_\gamma\big(z_t \mid z_{t-1}, a_{t-1}\big)\Big)\Big].\end{aligned} \tag{B.12}$$

For simplicity, the fixed conditioning variables $o_{1:t}$, z_0, and $a_{0:t-1}$ are omitted from the marginal distributions.

B.3 Factor-Wise KL Decomposition

For notational simplicity, throughout this subsection we simplify fixed $o_{1:t}$, z_0, and $a_{0:t-1}$ when no ambiguity can arise. Define

$$p_{\gamma_i}\big(z_{t,i}\big) = p_{\gamma_i}\big(z_{t,i} \mid z_{t-1,i}, a_{t-1}\big), q_{\phi_i}\big(z_{t,i}\big) = q_{\phi_i}\big(z_{t,i} \mid o_t, z_{t-1,i}.a_{t-1}\big).$$

By the factor-wise transition-independent dynamic in equation B.3, we now expand the KL divergence appearing in equation B.12. Using the factorization above, we have

$$\begin{aligned}
&D_{\mathrm{KL}}\Big(q_\phi\big(z_t \mid o_t, z_{t-1}, a_{t-1}\big) \,\Big\|\, p_\gamma\big(z_t \mid z_{t-1}, a_{t-1}\big)\Big) \\
&= D_{\mathrm{KL}}\Big(\prod_{j=1}^{K} q_{\phi_j}\big(z_{t,j}\big) \,\Big\|\, \prod_{i=1}^{K} p_{\gamma,i}\big(z_{t,i}\big)\Big) \\
&= \mathbb{E}_{\prod_{j=1}^{K} q_{\phi_j}(z_{t,j})}\left[\log \frac{\prod_{j=1}^{K} q_{\phi_j}\big(z_{t,j}\big)}{\prod_{i=1}^{K} p_{\gamma_i}\big(z_{t,i}\big)}\right] \\
&= \mathbb{E}_{\prod_{j=1}^{K} q_{\phi_j}(z_{t,j})}\left[\sum_{i=1}^{K} \log \frac{q_{\phi_i}\big(z_{t,i}\big)}{p_{\gamma_i}\big(z_{t,i}\big)}\right] \\
&= \sum_{i=1}^{K} \mathbb{E}_{\prod_{j=1}^{K} q_{\phi_j}(z_{t,j})}\left[\log \frac{q_{\phi_i}\big(z_{t,i}\big)}{p_{\gamma_i}\big(z_{t,i}\big)}\right] \\
&= \sum_{i=1}^{K} \mathbb{E}_{q_{\phi_i}(z_{t,i})}\left[\log \frac{q_{\phi_i}\big(z_{t,i}\big)}{p_{\gamma_i}\big(z_{t,i}\big)}\right] \\
&= \sum_{i=1}^{K} D_{\mathrm{KL}}\Big(q_{\phi_i}\big(z_{t,i}\big) \,\Big\|\, p_{\gamma,i}\big(z_{t,i}\big)\Big).
\end{aligned} \tag{B.13}$$

Moreover, the i-th KL term depends only on the previous latent state $z_{t-1,i}$. To make this dependence explicit, note

$$Q_{\phi_i,t-1}\big(z_{t-1,i}\big) = \int Q_{\phi,t-1}\big(z_{t-1}\big) \prod_{j\neq i} \mathrm{d}z_{t-1,j}, \tag{B.14}$$

as the edge distribution of the i-th factor of $Q_{\phi_i,t-1}$. Thus, we have

$$\mathbb{E}_{Q_{\phi,t-1}(z_{t-1})}\Big[D_{\mathrm{KL}}\Big(q_{\phi_i}\big(z_{t,i}\big) \,\Big\|\, p_{\gamma_i}\big(z_{t,i}\big)\Big)\Big] = \mathbb{E}_{Q_{\phi_i,t-1}(z_{t-1,i})}\Big[D_{\mathrm{KL}}\Big(q_{\phi_i}\big(z_{t,i}\big) \,\Big\|\, p_{\gamma_i}\big(z_{t,i}\big)\Big)\Big]. \tag{B.15}$$

B.4 TIF Training Loss

Substituting the factor-wise KL decomposition in equation B.13 and the factor-wise marginalization in equation B.15 into equation B.12 yields the factor-wise evidence lower bound used to train TIF:

$$\begin{aligned}
\mathcal{L}_{\mathrm{ELBO}} =& \sum_{t=1}^{T} \mathbb{E}_{Q_{\phi,t}(z_t)}\left[\log p_\theta\left(o_t \mid z_t\right)\right] \\
&- \sum_{t=1}^{T}\sum_{i=1}^{K} \mathbb{E}_{Q_{\phi_i,t-1}(z_{t-1,i})}\Big[D_{\mathrm{KL}}\Big(q_{\phi_i}\left(z_{t,i} \mid o_t, z_{t-1,i}, a_{t-1}\right) \,\Big\|\, p_{\gamma_i}\left(z_{t,i} \mid z_{t-1,i}, a_{t-1}\right)\Big)\Big].
\end{aligned} \tag{B.16}$$

C Environment Descriptions

C.1 Synthetic environments

We construct a synthetic environment whose transition dynamics satisfy the transition-independent conditional exponential-family structure. The true state consists of K scalar factors,

$$s_t = (s_{t,1}, \ldots, s_{t,K}) \in [-B, B]^K.$$

The initial state is sampled from a standard Gaussian distribution truncated to $[-B, B]^K$. At each time step, the action $a_t \in \{1, \ldots, |A|\}$ is sampled uniformly and independently of the current state. Given s_t and a_t, the next state is generated according to

$$p(s_{t+1} \mid s_t, a_t) = \prod_{i=1}^{K} p(s_{t+1,i} \mid s_{t,i}, a_t).$$

Each factor transition is a one-dimensional conditional exponential-family distribution supported on $[-B, B]$:

$$p(u \mid v, a) = \frac{\exp\left[\boldsymbol{\phi}(u)\boldsymbol{\lambda}_i(v, a)\right]}{\int_{-B}^{B} \exp\left[\boldsymbol{\phi}(\tilde{u})\boldsymbol{\lambda}_i(v, a)\right] \, \mathrm{d}\tilde{u}},$$

where the sufficient statistic is defined by

$$\boldsymbol{\phi}(u) = u + qu^2$$

The corresponding parameters are defined as

$$\lambda_i(v, a) = L \tanh\left(\frac{\tau}{L}\left[b_1 v + b_2 \tanh(v) + b_3 \sin(v) + \rho \boldsymbol{H}_{a,i} v\right]\right),$$

where $\boldsymbol{H} \in \mathbb{R}^{M \times K}$ is the action-response encoding matrix with columns centered and orthogonalized such that

$$\boldsymbol{H}^\top \boldsymbol{H} = M \boldsymbol{I}_K, \qquad \sum_{a=1}^{M} H_{a,i} = 0.$$

Therefore, different actions induce distinct transition responses across the state factors. The outer hyperbolic tangent is used only to constrain the range of the corresponding parameters. The normalization constants and the sampling process for the transition distributions are implemented using one-dimensional numerical integration.

The high-dimensional observations are generated by an invertible nonlinear mixing function. First, each state factor is transformed independently:

$$u_t = s_t + c \sin(s_t)$$

The transformed state is then linearly mixed using an invertible matrix $\boldsymbol{A} \in \mathbb{R}^{K \times K}$:

$$o_t = \boldsymbol{A} u_t.$$

The derivative of the scalar transformation is strictly positive when $0 < c < 1$. Consequently, the component-wise transformation is strictly monotone. Together with the invertibility of A, this implies that the overall observation mapping is invertible.

C.2 MiniGrid

To evaluate TIF from high-dimensional visual observations, we construct a set of visual grid-world environments based on MiniGrid. To make the semantics of the ground-truth factors explicit, we remove the agent's orientation variable from the standard MiniGrid environments. The agent is rendered as an orientation-free square and,therefore, its ground-truth state does not include orientation. Accordingly, we replace the original orientation-dependent turning and forward actions with actions that directly move the agent along the absolute grid directions.

During training, the model receives only $64 \times 64 \times 3$ RGB observations o_t and actions a_t. The ground-truth state variables are used only to evaluate factor identifiability after training and are not used for model training or model selection.

Variant of MiniGrid Empty. We construct three environment variants based on MiniGrid Empty. The active region of the agent is a 7×7 grid, and the position of the agent is represented by$s_t = (x_t^a, y_t^a)$. The 4-action environment contains no visible goal. Its action space consists of upward, downward, leftward, and rightward movements. Each transition changes at most one coordinate of the agent position.

The 8-action environment additionally includes the four diagonal movements. Thus, a single action can change both coordinates of the agent position. This setting is used to test whether factor recovery relies on the restrictive assumption that each transition changes at most one factor.

The random-goal environment extends the 4-action environment with a visible goal. The goal position is sampled at the beginning of each episode and remains fixed throughout the episode. The ground-truth state is therefore $s_t = (x_t^a, y_t^a, x^g, y^g)$, where (x_t^a, y_t^a) denotes the agent position and (x^g, y^g) denotes the goal position. Because the goal position is sampled at the beginning of each episode and is unaffected by actions, the goal coordinates are uncontrollable factors.

Variant of MiniGrid FourRooms. We construct a fixed-layout environment based on MiniGrid FourRooms. The feasible region of the agent is a 9×9 grid, and internal walls divide the grid into four rooms. The adjacent rooms are connected through fixed doors.

At the beginning of each episode, the agent is initialized in a random position. The image shows a goal in a fixed position, and the episode ends when the agent reaches the goal. The action space consists of four movements along the absolute directions of the grid: up, down, left, and right. If an action points toward a wall or the boundary of the grid, the agent remains at its current position.

Because the room layout and the goal position are fixed, the ground-truth state contains only the agent position: $s_t = (x_t^a, y_t^a)$. Compared with the Empty environments, the internal walls make the outcome of an action depend jointly on the agent's current two-dimensional position. This environment is therefore used to evaluate factor representations and policy learning under more complex navigation dynamics.

D Experimental Details

D.1 Baselines and Metrics

Identifiability baselines. In the factor identifiability experiments, we compare TIF with three representative categories of disentanglement methods. The first category consists of static unsupervised disentanglement methods, including β-VAE and FactorVAE. The second category consists of representation learning methods that exploit temporal continuity, including SlowVAE and PCL. SlowVAE assumes that only one factor changes in each transition. PCL learns representations by distinguishing pairs of consecutive observations from pairs formed by randomly shuffling observations. The third category consists of identifiable representation learning methods that use auxiliary information or changing mechanisms, including GCL, iVAE, and TDRL. GCL uses $u_t = (o_{t-1}, a_{t-1})$ as the auxiliary variable. iVAE uses a diagonal Gaussian prior conditioned on (o_{t-1}, a_{t-1}). TDRL uses an action embedding as the context for dynamical changes.

Policy-learning baselines. In the policy-learning experiments, we compare TIF with DreamerV2 and several factorized world model methods, including IFactor, DisWM, and Iso-Dream. For DreamerV2, we consider both the capacity-matched version and the DreamerV2 (Large) version. For DisWM, we do not use an additional pretraining stage. All methods use the same environment interaction budget and policy evaluation protocol.

Metrics. We use R^2 and the mean correlation coefficient (MCC) to evaluate the correspondence between the learned representations and the ground-truth factors. For R^2, we separately fit factor-wise mappings from z to s and from s to z using RBF kernel ridge regression. We report the average bidirectional R^2 after matching the factors, denoted by Bi-R^2, as well as the strongest R^2 between unmatched factors, denoted by Off-diagonal R^2. MCC is computed from the absolute Spearman correlation matrix between the learned factors and the ground-truth factors, followed by optimal one-to-one matching. All results are evaluated on independent test set and reported as the mean and standard deviation over five random seeds.

Table 4: Architecture of the factorized world model used in the MiniGrid experiments.

Component	Configuration	Dimension
Shared observation encoder	Input RGB observation	$64 \times 64 \times 3$
	4 convolutional layers with channels $(48, 96, 192, 384)$, kernel size 4, stride 2, and ELU activations	–
	Flattened observation embedding e_t	1536
Each Factor-RSSM branch	Stochastic state $z_{t,i}$	1
	Deterministic state $h_{t,i}$	32
	Transition hidden layer	128
		$z_{t-1,i}, h_{t-1,i}$
	Prior input	a_{t-1}
	Prior output: Gaussian mean and standard deviation	2
	Posterior input	$h_{t,i}, e_t$
	Posterior output: Gaussian mean and standard deviation	2
Shared observation decoder	Concatenated states from K branches	$33K$
	Dense layer followed by reshaping	$1536 \times 1 \times 1$
	4 transposed convolutional layers with channels $(192, 96, 48, 3)$, kernel sizes $(5, 5, 6, 6)$, stride 2, and ELU activations except at the output layer	–
	Reconstructed RGB observation	$64 \times 64 \times 3$

D.2 Model Architectures

Synthetic Environments. We use an explicit-likelihood implementation of TIF. The encoder is a RealNVP consisting of 8 affine coupling layers. It maps each observation o_t to a bounded latent state $z_t \in [-B, B]^K$, and the inverse mapping defines the corresponding observation model.

The latent transition model follows the factor-wise decomposition. Each factor transition is modeled by a one-dimensional conditional exponential family. The sufficient statistics are parameterized by a strictly monotone scalar neural network, while the corresponding parameter network for factor i receives only the corresponding previous latent factor $z_{t,i}$ and the action a_{t-1}. The coupling networks in RealNVP and the corresponding parameter networks are three-layer MLPs with hidden width 128. The normalizing constants of the transition distributions are approximated using Gauss-Legendre quadrature with 128 nodes.

MiniGrid environments. We use the same factorized world-model architecture. The only architectural choice specific to the environment is the number of recurrent state space model (RSSM) branches K, which is set according to the number of ground-truth factors in the environment. The model applies a shared convolutional encoder to map each $64 \times 64 \times 3$ RGB observation to a visual embedding. The latent are then modeled by K independent continuous Gaussian RSSM branches. Each branch contains a one-dimensional stochastic state $z_{t,i}$, a 32-dimensional deterministic state $h_{t,i}$, and a transition hidden layer with width 128. For branch i, the transition prior is conditioned only on the previous stochastic state $z_{t-1,i}$, the previous deterministic state $h_{t-1,i}$, and the action a_{t-1}. In contrast, the posterior representation additionally receives the shared observation embedding. The stochastic and deterministic states from all K branches are concatenated and provided to a shared observation decoder, which reconstructs the current RGB observation. The complete architecture and the dimensionality of each component are summarized in Table 4.

To ensure a fair comparison, all baseline methods use the same observation encoder, observation decoder, and total latent-state dimensionality as TIF. For methods that use a joint dynamics model rather than independent factor branches, the stochastic-state dimension, the deterministic-state dimension, and the hidden transition dimension are set to K, $32K$, and $128K$, respectively. Thus, these dimensions match the aggregate dimensions of the K TIF branches. In addition, we report DreamerV2 (large), whose corresponding dimensions are set to 32, 512 and 512 respectively.

D.3 Training and Hyperparameters

Synthetic environments. All experiments use $K = 5$ ground-truth factors and $|A| = 8$ actions, with state support $[-5, 5]^5$. Each trajectory contains five state transitions, corresponding to six consecutive states and observations. The training set contains 8,000 trajectories, corresponding to 40k environment steps, and the test set contains 2,000 independent trajectories. At each time

Table 5: Training configurations for the MiniGrid environments.

Environment	Interaction steps	Training steps	Sequence length
Empty 4-action	40k	10k	11
Empty 8-action	40k	10k	11
Random-goal	40k	20k	11
Four-Rooms	200k	20k	26

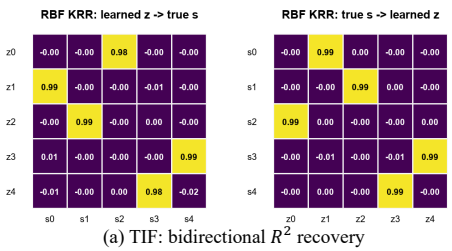
step, the action is sampled uniformly and independently from the eight discrete actions. We set $q = 0.05$, $\tau = 0.35$, $(b_1, b_2, b_3) = (0.20, 0.25, 0.10)$, $\rho = 0.50$, $L = 4.0$, and $c = 0.35$. The latent dimensionality of all methods is set to the number of ground-truth factors K. We train all models using the Adam optimizer for 100 epochs with a batch size of 512 and a learning rate of 10^{-3}.

MiniGrid environments. The Empty 4-action, Empty 8-action, and FourRooms environments use $K = 2$ factor branches, corresponding to the two position factors of the agent. The Random-goal environment uses $K = 4$ factor branches, corresponding to the four position factors of the agent and the goal. All MiniGrid models are trained using the Adam optimizer with a batch size of 32 and a learning rate of 10^{-4}. The KL scale and KL balance coefficient are set to 1.0 and 0.5, respectively. The number of interaction steps, training steps, and sequence length used for each environment are reported in Table 5.

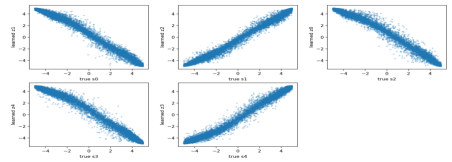
E Empirical Results

E.1 Synthetic environments

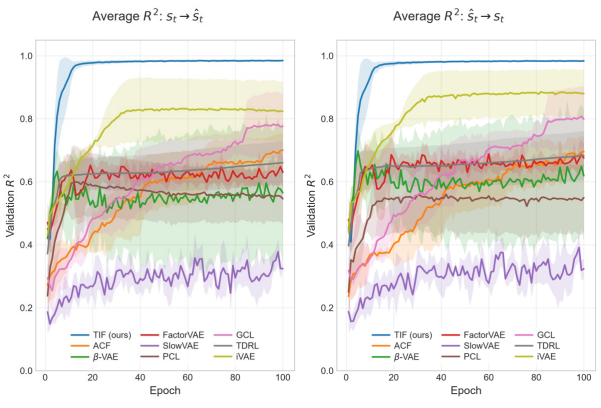
This section presents the complete factor recovery results in the synthetic environment, including detailed visualizations of TIF's factor recovery, the training process (Figure 8), and the bidirectional R^2 matrices for the baseline methods (Figure 9).



(a) TIF: bidirectional R^2 recovery



(b) TIF: matched factor relationships



(c) Training: bidirectional R^2 recovery

Figure 8: Detailed factor recovery results in the synthetic environment. (a) Bidirectional R^2 matrices between the learned factors and the ground-truth factors for TIF. (b) Matched factor-wise relationships recovered by TIF. (c) Training curves of the average bidirectional R^2 for TIF and the comparison methods.

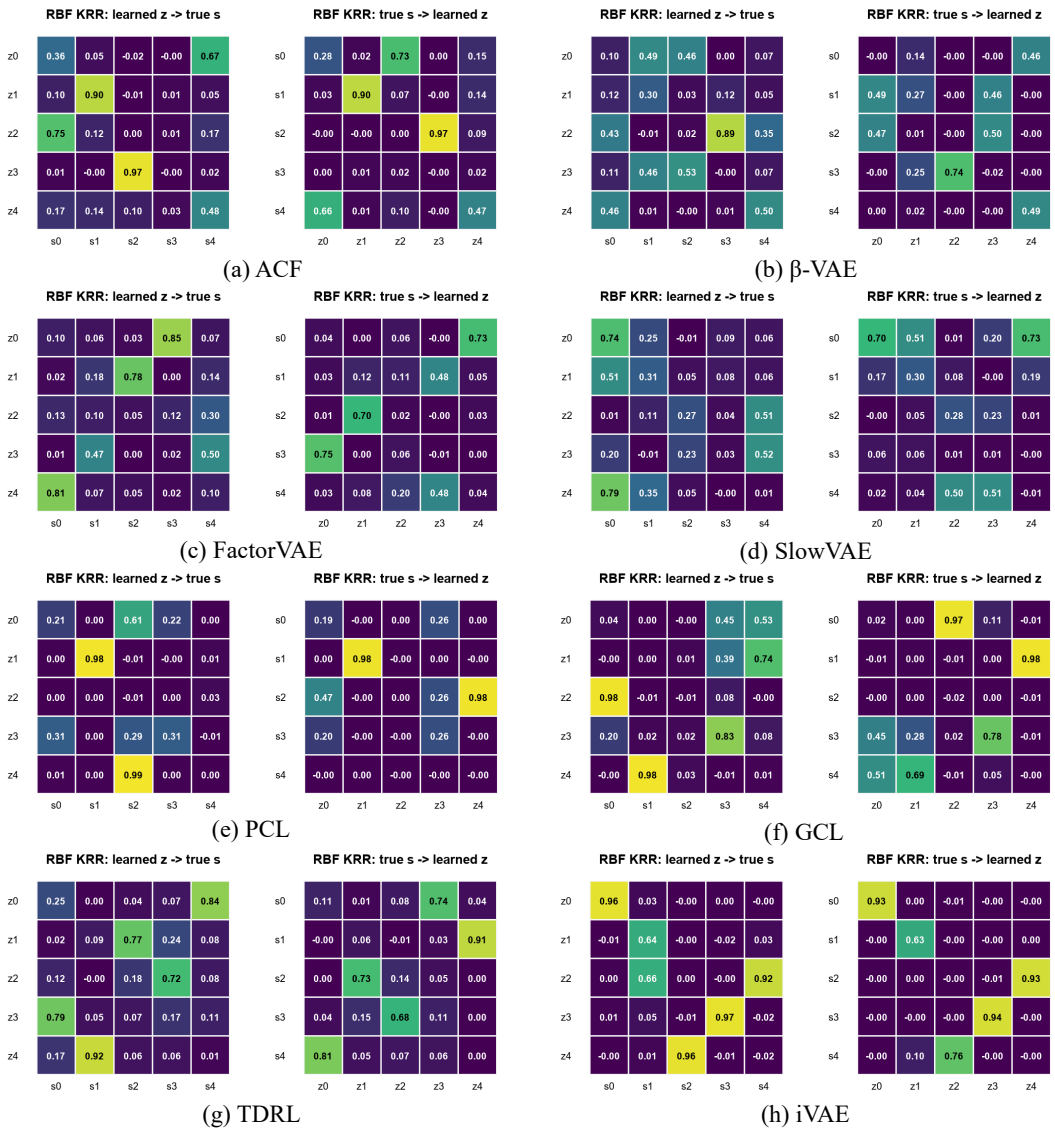


Figure 9: Bidirectional factor recovery of the baseline methods in the synthetic environment. Each pair of matrices reports the RBF-KRR R^2 from the learned factors to the ground-truth state variables and from the ground-truth state variables to the learned factors, respectively.

E.2 MiniGrid Environment

Figure 10 presents the SlowVAE factor recovery results in the 4-action and 8-action environments. Similar to the ACF results shown in Figure 3, SlowVAE forms relatively clear factor correspondences in the 4-action environment, but exhibits substantial cross-factor mixing in the 8-action environment. This phenomenon is related to the identifiability mechanism on which the two methods rely: ACF relies on sparse action-to-factor effects and distinguishes different state variables through the sparse relationship between actions and factors. While SlowVAE is based on temporal sparsity of factor changes, assuming that only a small number of factors change in each one-step transition. When diagonal actions can change x and y simultaneously, this sparsity structure weakens, making the two state factors difficult to separate through this mechanism.

In contrast, TIF does not rely on sparse action-to-factor effects and does not require one-step state changes to be sparse. Instead, it uses factor-wise transition independence and action-response diversity. Therefore, it maintains stable factor correspondences under both action settings.

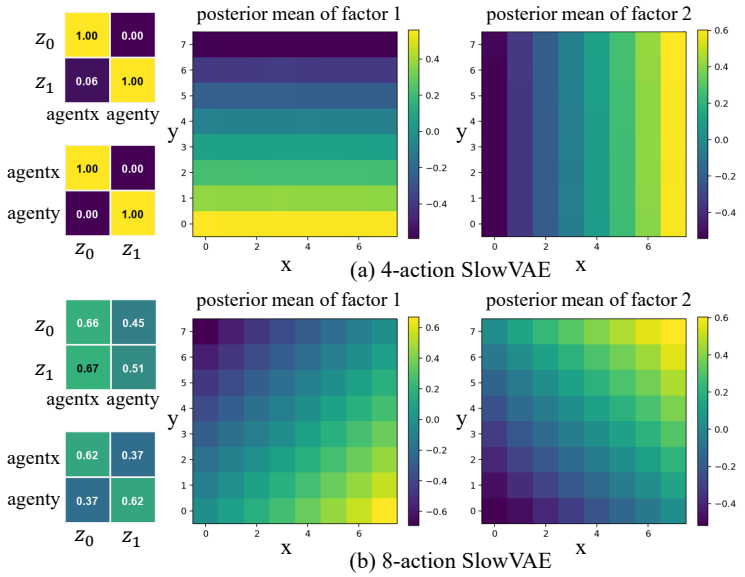
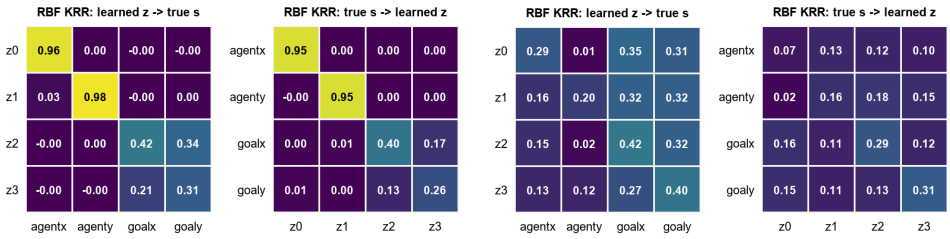


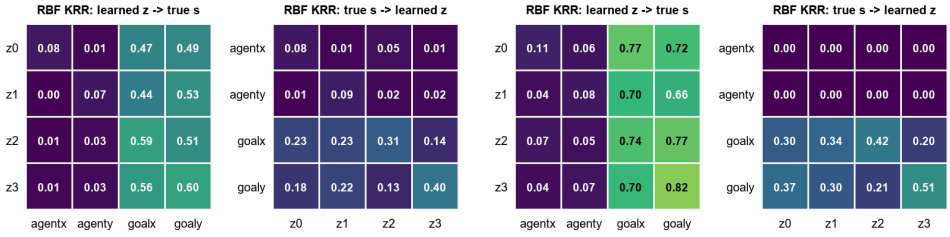
Figure 10: Factor recovery of SlowVAE in the 4-action and 8-action environments.

Figure 11 shows the factor recovery results in the random-goal environment. For TIF, the two learned factors consistently recover the agent's x- and y-coordinates, respectively. Since the goal position is fixed within each episode and is not controlled by actions, its x- and y-coordinates cannot be identified at the factor level from action responses. Therefore, they do not form one-to-one correspondences with individual learned factors. Nevertheless, the remaining latent factors retain clear information about the goal position, but encode it in a mixed form. This is consistent with our theoretical expectation: TIF's identifiability guarantee applies to controllable factors. The representation model may retain uncontrollable variables, but their factor recovery is not guaranteed.

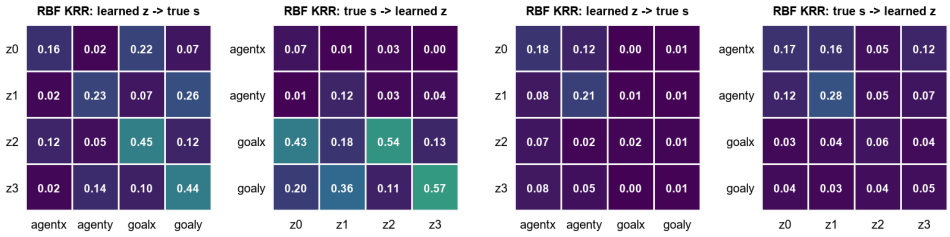
Interestingly, ACF exhibits different behavior. As shown in Figure 12, when the latent dimension is set to 2, ACF can recover the agent's x- and y-coordinates, but almost no information about the goal position is retained. When the latent dimension is increased to 4, the additional latent dimensions do not represent the goal position; instead, they redundantly encode the agent's x- and y-coordinates, resulting in redundant controllable representations. This is consistent with ACF's reliance on controllability to distinguish the underlying mechanisms of variables. In contrast, the additional representation capacity of TIF can absorb uncontrollable environmental variations while preserving a stable one-to-one correspondence for the controllable factors, indicating better adaptability when the number of controllable factors is smaller than the dimensionality of the full state.



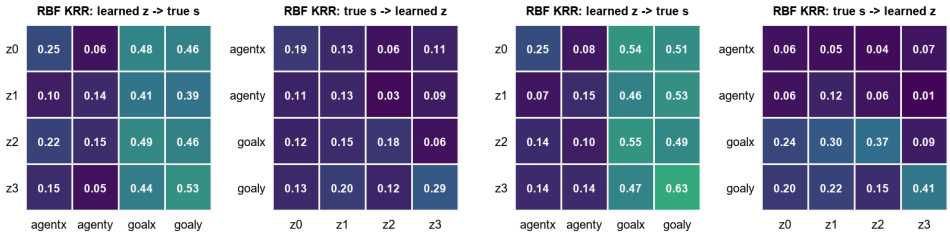
(a) TIF (b) β-VAE



(c) FactorVAE (d) SlowVAE

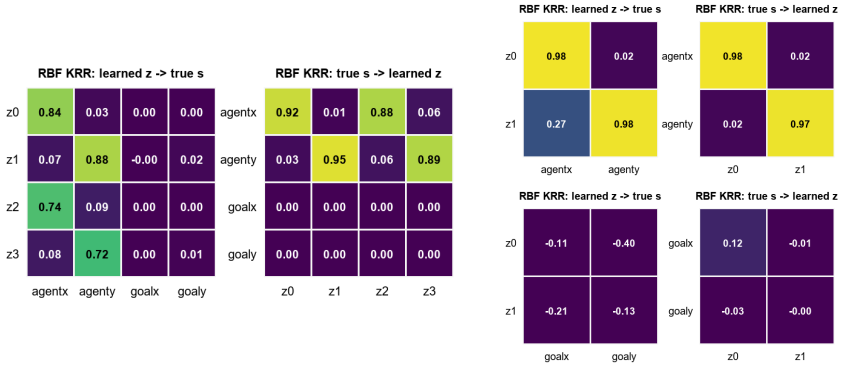


(e) PCL (f) GCL



(g) TDRL (h) iVAE

Figure 11: Bidirectional factor recovery of TIF and baselines in the random-goal setting.



(a) ACF latent=4 (b) ACF latent=2

Figure 12: Factor recovery of ACF with different latent dimensions in the random-goal setting.