CACTI: CROSS-ACTION COUNTERFACTUAL TRANSITION INFERENCE FROM OFFLINE DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Same-event counterfactual transition inference asks what the next state would have been under a different action while keeping the disturbance realized in the observed event fixed. Accurately fitting action-conditional transition distributions alone does not determine these same-event pairings, especially under uneven action coverage. We propose CACTI, a structured transition causal model that combines a shared state-dependent action response with a conditional invertible model of event-specific variation. This structure lets variation across related actions constrain queries with little direct data while retaining information from the factual event. Under a linear shared-response model and sequential exogeneity, we prove that, at a fixed state and without additional cross-state restrictions, the response to a query action difference is identifiable if and only if that direction lies in the range of the conditional action covariance. We further derive a local counterfactual error bound in terms of shared-response mismatch and action distance. We evaluate CACTI against same-event simulator oracles on four control tasks. It achieves the lowest mean counterfactual error on both all nonfactual queries and low-direct-support subsets, outperforming the compared predictive and explicit-abduction baselines. Policy-learning experiments with limited data further show that the generated counterfactual transitions can provide useful signals for downstream training.

1 INTRODUCTION

Same-event counterfactual transition inference asks a specific question: given an observed transition, what would the next state have been under a different action while preserving the disturbance realized in that event? This is different from predicting a plausible outcome under the new action. Consider a vehicle that turns right under a particular crosswind and drifts off course. We ask what would have happened had the vehicle turned left under the same gust. Two models can fit the observed transition equally well yet attribute different parts of the motion to steering and wind. They may therefore agree on the action-conditional outcome distributions while giving different counterfactual answers. Accurate prediction alone does not determine which outcomes across actions correspond to the same event (Nasr-Esfahany & Kiciman, 2023).

Prior work uses causal structure to generate counterfactual data (Pitis et al., 2020; 2022) or explicitly abducts and reuses exogenous variables across interventions (Oberst & Sontag, 2019; Lu et al., 2020; Tsirtsis & Rodriguez, 2023). We focus on a related difficulty that becomes important in offline reinforcement learning: action coverage is often uneven (Kumar et al., 2020; Yu et al., 2020). The queried action may have little direct support, while nearby or directionally related actions remain well represented in the data. Such a query is poorly observed, but not necessarily unconstrained. This motivates a shared response model that lets action variation observed elsewhere inform the queried action. We refer to this regime as low direct support.

We introduce CACTI, a structured transition SCM for same-event counterfactual inference:

$$\Delta S_{t+1} = m_\theta(S_t) + B_\theta(S_t)A_t + g_\psi(S_t, U_{t+1}). \tag{1}$$

The state term m_θ captures predictable evolution from the current state. The shared response B_θ captures how actions systematically change the transition. The exogenous term g_ψ captures event-specific variation and does not directly receive the action. This separation lets different actions share

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.

response information while keeping the event-specific contribution tied to the factual transition. We model the exogenous term with a state-conditioned invertible flow, which gives a tractable factual likelihood and allows all components to be learned from factual transitions alone. No true disturbances or counterfactual labels are required.

Our analysis assumes sequential exogeneity: conditioned on the recorded state, the behavior action is independent of the next-step disturbance. CACTI allows all components to vary nonlinearly with the state, but it restricts the action response at a fixed state and rules out unrestricted action-disturbance interactions. These restrictions are what make cross-action sharing possible. They also define the scope of the model.

We then ask which counterfactual queries the offline data can actually support. Under the shared-response model and sequential exogeneity, the response to an action difference $d = a' - a$ is identified from the observational distribution if and only if $d \in \mathrm{Range}(\Sigma_A(s))$, where $\Sigma_A(s)$ is the conditional action covariance. The key point is simple: little direct data for the queried action does not imply that the queried direction is unidentifiable. What matters is whether the required action direction is represented in the observed variation. Directional coverage, however, does not guarantee finite-sample stability or accuracy for large interventions. We therefore also decompose counterfactual error into event-matching and response-estimation terms, and bound the additional error caused by local response mismatch and nonlinear action effects.

We evaluate counterfactual fidelity separately from downstream policy utility. Across MountainCar11, PointMaze, Reacher, and Shadow Hand RotateZ, we use a simulator oracle that restores the factual state, reuses the same realized disturbance, and changes only the action. The oracle disturbance is used only to construct evaluation targets and is never given to the learned models. CACTI achieves lower mean counterfactual error than the compared predictive and explicit-abduction baselines on both all nonfactual queries and the low-direct-support subsets in all four environments. Additional controls probe the role of factual event information. We then test the generated transitions in low-data policy learning on two environments, treating policy performance as a downstream application rather than a substitute for counterfactual evaluation.

Our main contributions are:

1. A structured model for same-event counterfactual transitions. CACTI combines a shared action response with a conditional invertible exogenous model, allowing factual transitions to jointly constrain systematic action effects and event-specific variation.
2. Query-level identifiability and local error analysis. We characterize which action differences are identifiable under the shared-response model and separate errors due to event mismatch, response mismatch, and nonlinear action effects.
3. Direct counterfactual evaluation. We evaluate same-event transition fidelity across four control tasks, including low-direct-support queries and event-matching diagnostics, and further test the resulting transitions for low-data policy learning.

2 Problem Setup and Counterfactual Target

Data and setting. We consider an offline transition dataset $\mathcal{D} = \{(s_i, a_i, s_i^+)\}_{i=1}^N$, where each sample records a state, the action taken, and the resulting next state. We model the true transition with an unknown structural causal model (Peters et al., 2017),

$$S_{t+1} = f^\star(S_t, A_t, U_{t+1}), \tag{2}$$

where U_{t+1} captures event-specific factors not explicitly represented by the recorded state and action. Training uses factual transitions only. Neither the realized U_{t+1} nor counterfactual outcomes are observed. We assume sequential exogeneity, $U_{t+1} \perp A_t \mid S_t$, so that, conditioned on the recorded state, the behavior action does not anticipate the disturbance realized in the next transition.

Same-event counterfactual query. Consider a factual transition $e_t = (s_t, a_t, s_{t+1})$, with realized exogenous variable u_{t+1}. For an alternative action a'_t, the target is

$$s^\star_{t+1}(a'_t) = f^\star(s_t, a'_t, u_{t+1}). \tag{3}$$

The query changes only the action while holding the factual state and the same exogenous realization fixed. It therefore asks about the same event, rather than a newly sampled transition under a'_t. The model receives the full factual transition e_t and the query action a'_t, and predicts $\widehat{s}_{t+1}(a'_t; e_t)$. Thus, the factual next state is part of the query input, while the realized u_{t+1} and the counterfactual target remain unobserved.

Structured learning. Matching action-conditional transition distributions is not enough to determine a same-event counterfactual. Different SCMs can induce the same observed distributions while pairing outcomes across actions differently. A model must therefore specify how outcomes are linked across action worlds, not only how each action predicts the next state (Nasr-Esfahany et al., 2023). We address this with a structured transition model. The next section introduces CACTI, its training objective, and its counterfactual query rule. We then analyze which queries are identifiable under this structure and how local model mismatch affects counterfactual error.

3 CACTI: Separating Action Effects from Factual Disturbances

3.1 Shared-Response Transition Model

CACTI models one-step state changes as

$$\Delta S_{t+1} = m_\theta(S_t) + B_\theta(S_t)A_t + g_\psi(S_t, U_{t+1}), \tag{4}$$

where $\Delta S_{t+1} = S_{t+1} - S_t$. The three terms play different roles. $m_\theta(S_t)$ captures evolution explained by the current state, $B_\theta(S_t)A_t$ captures the systematic effect of the action, and $g_\psi(S_t, U_{t+1})$ captures event-specific variation. In the vehicle example from the introduction, m_θ describes motion implied by the current speed and heading, $B_\theta A_t$ describes the steering response, and g_ψ captures unrecorded influences such as the particular crosswind realized in that event. For convenience, we write

$$\mu_\theta(s, a) := m_\theta(s) + B_\theta(s)a. \tag{5}$$

The goal is to separate response patterns that can be learned across actions from variation specific to the observed event.

Shared action response. At a fixed state s, all actions share the same state-dependent response $B_\theta(s)$. In the vehicle example, different steering magnitudes provide evidence about the same steering response at a given speed and heading. A rarely observed query action can therefore still benefit from variation seen under other actions. CACTI pools this cross-action evidence instead of treating each action as an unrelated prediction problem. Because $B_\theta(s)$ depends on the state, the response can still change with the operating condition; the model does not assume one global linear dynamics law.

State-dependent event variation. The exogenous term $g_\psi(s, u)$ models variation that is specific to a realized event. It depends on the state because the same disturbance can have different effects under different system conditions. The same crosswind, for example, may cause a different displacement at different speeds or headings.

At a fixed state, all actions share the same exogenous map. An intervention therefore changes the action response without changing the coordinate system used to represent the factual event. Action-specific latent relabelings can preserve action-conditional outcome distributions while changing same-event pairings. Sharing the exogenous map rules out this specific degree of freedom. Appendix B.1 gives the formal construction.

We implement g_ψ with a state-conditioned invertible flow and assume that, for each fixed s, $g_{\psi,s} : u \mapsto g_\psi(s, u)$ is invertible. This choice serves two purposes. First, it gives CACTI a tractable conditional density for state-dependent event variation, so the model can be trained directly from factual transitions by likelihood. Second, it maps an observed factual transition to an internal event representation that can be reused in a counterfactual query (Pawlowski et al., 2020; Khemakhem et al., 2021). The latent coordinate U belongs to the learned model; it need not use the same numerical coordinates as the underlying physical disturbance (Chen & Du, 2025).

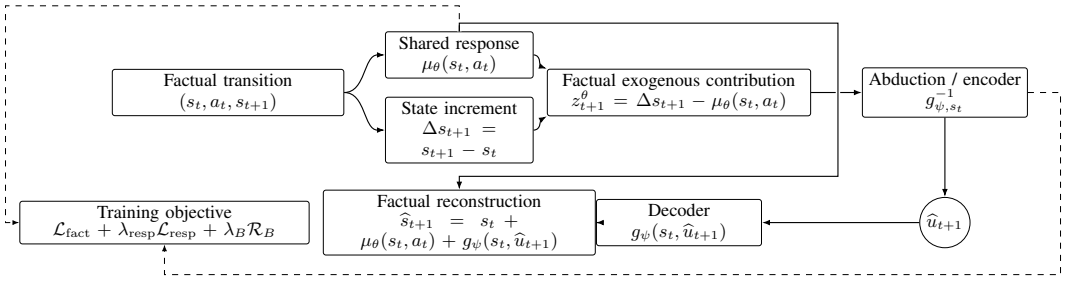

Figure 1: CACTI training pipeline. The shared response models systematic action effects, while the event-specific contribution is encoded as $\widehat{u}_{t+1}$ and decoded for factual reconstruction. Dashed arrows denote training constraints.

CACTI therefore shares structure in two ways: $B_\theta(s)$ pools action-response information across actions, while $g_{\psi,s}$ provides a common event representation across those actions at the same state. The state dependence of m_θ, B_θ, and g_ψ remains flexible, while action-dependent variation is concentrated in the shared response. The next section shows how these components are learned jointly from factual transitions.

3.2 Learning from Factual Transitions

CACTI learns the shared action response and event variation jointly from factual transitions. For a transition (s, a, s^+), let

$$\Delta s = s^+ - s, \qquad z_\theta = \Delta s - \mu_\theta(s, a). \tag{6}$$

The residual z_θ contains the variation not explained by the systematic state-action response. We model its state-dependent distribution with the conditional flow $g_{\psi,s}$. Invertibility gives a tractable conditional density $p_{\theta,\psi}(\Delta S \mid S, A)$, which we fit with

$$\mathcal{L}_{\text{fact}} = -\mathbb{E}_{(s,a,s^+)\sim\mathcal{D}}\left[\log p_{\theta,\psi}(\Delta s \mid s, a)\right]. \tag{7}$$

This likelihood trains the full transition distribution from factual data alone. No true disturbance or counterfactual target is observed.

The likelihood couples the shared response and the event model, but it does not give the response branch a direct training signal. We therefore add

$$\mathcal{L}_{\text{resp}} = \mathbb{E}_{\mathcal{D}}\left[\|\Delta s - m_\theta(s) - B_\theta(s)a\|_2^2\right], \tag{8}$$

which directly trains m_θ and B_θ from observed state-action variation. We also regularize the response magnitude with

$$R_B = \mathbb{E}_{\mathcal{D}}\left[\|B_\theta(s)\|_F^2\right]. \tag{9}$$

The full objective is

$$\mathcal{L}_{\text{CACTI}} = \mathcal{L}_{\text{fact}} + \lambda_{\text{resp}}\mathcal{L}_{\text{resp}} + \lambda_B R_B. \tag{10}$$

We first stabilize the response branch and then jointly optimize m_θ, B_θ, and g_ψ. Architecture and optimization details are given in Appendices D.1 and A.2.

Figure 1 summarizes the corresponding learning pipeline.

3.3 Same-Event Counterfactual Query

Given a factual transition $e_t = (s_t, a_t, s_{t+1})$ and a query action a'_t, CACTI first recovers the event representation implied by the factual transition:

$$z_{t+1} = (s_{t+1} - s_t) - \mu_\theta(s_t, a_t), \qquad \widehat{u}_{t+1} = g^{-1}_{\psi,s_t}(z_{t+1}). \tag{11}$$

The counterfactual changes only the action while keeping the same event representation:

$$\widehat{s}_{t+1}(a'_t) = s_t + \mu_\theta(s_t, a'_t) + g_\psi(s_t, \widehat{u}_{t+1}). \tag{12}$$

The factual transition fixes the event-specific contribution; the query action changes the action response.

Algorithm 1 CACTI Training and Same-Event Inference

Require: Offline transitions $\mathcal{D}$, iterations K

$\mu_\theta(s,a) \leftarrow m_\theta(s) + B_\theta(s)a$

Training

for $k = 1, \ldots, K$ **do**

  Sample $\mathcal{B} = \{(s, a, s^+)\} \sim \mathcal{D}$

  $\Delta s \leftarrow s^+ - s, \quad z^\theta \leftarrow \Delta s - \mu_\theta(s,a)$

  $\mathcal{L} \leftarrow \mathcal{L}_{\text{fact}} + \lambda_{\text{resp}}\mathcal{L}_{\text{resp}} + \lambda_B\mathcal{R}_B$

  Update (θ, ψ) by minimizing $\mathcal{L}$

end for

Same-event query

Input: factual transition (s_t, a_t, s_{t+1}), query action a'_t

$z_{t+1} \leftarrow (s_{t+1} - s_t) - \mu_\theta(s_t, a_t)$ ▷ Factual residual

$\widehat{u}_{t+1} \leftarrow g^{-1}_{\psi,s_t}(z_{t+1})$ ▷ Model-space event representation

$\widehat{s}_{t+1}(a'_t) \leftarrow s_t + \mu_\theta(s_t, a'_t) + g_\psi(s_t, \widehat{u}_{t+1})$

Equivalent form: $\widehat{s}_{t+1}(a'_t) = s_{t+1} + \mu_\theta(s_t, a'_t) - \mu_\theta(s_t, a_t)$ ▷ Exact inversion

return $\widehat{s}_{t+1}(a'_t)$

Under exact inversion, the same query admits an equivalent residual form:

$$\begin{aligned}\widehat{s}_{t+1}(a'_t) &= s_{t+1} + \mu_\theta(s_t, a'_t) - \mu_\theta(s_t, a_t) \\ &= s_{t+1} + B_\theta(s_t)(a'_t - a_t).\end{aligned} \tag{13}$$

Thus, the factual transition carries the event-specific contribution, while the shared response determines how the outcome changes with the action. The conditional flow models the factual transition distribution during learning; the one-step counterfactual query depends on the learned action contrast.

This reduces the remaining question to the action difference $d_t = a'_t - a_t$. When does the offline data support the response $B_\theta(s_t)d_t$, and how much error arises when the shared response deviates from the true action effect? The next section studies these two questions through action-direction coverage and local response error.

4 Identifiability and Error of Same-Event Queries

Section 3.3 shows that a one-step same-event query reduces to preserving the factual event contribution and estimating the response to the action difference $d_t = a'_t - a_t$. This leaves two questions. When does the action variation in the offline data identify the target response? How accurate is the shared-response approximation when it does not exactly capture the true action effect? Section 4.1 answers the first through action-direction coverage. Section 4.2 separates event mismatch from response mismatch and bounds the local counterfactual error. Proofs are given in Appendix B.

4.1 Which Action Responses Are Identifiable from Offline Data?

Section 3.3 reduces a one-step same-event query to the response along the action difference $d_t = a'_t - a_t$. This raises a basic question: when does the action variation in the offline data determine that response?

To isolate this information question, we first consider the case where the shared-response model is locally exact:

$$\Delta S_{t+1} = m^\star(S_t) + B^\star(S_t)A_t + g^\star(S_t, U_{t+1}), \tag{14}$$

with sequential exogeneity

$$U_{t+1} \perp A_t \mid S_t. \tag{15}$$

At a fixed state s, define

$$\Sigma_A(s) = \mathrm{Cov}(A_t \mid S_t = s), \qquad C_{\Delta A}(s) = \mathrm{Cov}(\Delta S_{t+1}, A_t \mid S_t = s). \tag{16}$$

The range of $\Sigma_A(s)$ is the span of action directions that vary in the offline data.

Theorem 1 (Action-direction identifiability) *Assume the relevant conditional second moments exist, and do not impose additional cross-state restrictions. For any action difference d, the response $B^\star(s)d$ is uniquely determined by the observational distribution if and only if*

$$d \in \mathrm{Range}(\Sigma_A(s)). \tag{17}$$

When this condition holds,

$$B^\star(s)d = C_{\Delta A}(s)\Sigma_A(s)^\dagger d. \tag{18}$$

The result follows from

$$C_{\Delta A}(s) = B^\star(s)\Sigma_A(s), \tag{19}$$

which is implied by sequential exogeneity. The data therefore constrain the response only along action directions that actually vary. If d lies in $\mathrm{Range}(\Sigma_A(s))$, every model consistent with the observational distribution agrees on the response along d. If d has a nonzero component in $\mathrm{Null}(\Sigma_A(s))$, one can change the shared response along that component and compensate the state-only term so that the full conditional observational distribution at s remains unchanged while the query effect changes. The full proof is given in Appendix B.2.

What matters is not how often the query action appears, but whether the required action direction is represented in the data. Suppose a vehicle has both steering and acceleration controls, but only steering varies in the offline dataset. The local steering response can then be identified, while the acceleration response cannot. Conversely, a query action need not have been observed exactly: its response can still be identified if its difference from the factual action lies in the span of observed action variation.

Thus, low direct support does not necessarily imply a lack of counterfactual information. Directional coverage is the key identifiability condition in this model. Identifiability, however, does not guarantee accurate finite-sample estimation or reliable extrapolation over large action changes. The next section studies the error introduced when the shared response only approximates the true action effect.

4.2 Where Does Counterfactual Error Come From?

Section 4.1 asked when the offline data contain enough action variation to identify a query response. We now consider the second source of difficulty: even when the query direction is supported, the learned shared response may differ from the true same-event action effect.

For a factual event $e_t = (s_t, a_t, s_{t+1})$ and query action a'_t, let $d_t = a'_t - a_t$ and define the true same-event action contrast

$$\tau_t^\star(a'_t) = F^\star(s_t, a'_t, u_{t+1}) - F^\star(s_t, a_t, u_{t+1}). \tag{20}$$

Since $F^\star(s_t, a_t, u_{t+1}) = s_{t+1}$, the true counterfactual is

$$s^\star_{t+1}(a'_t) = s_{t+1} + \tau_t^\star(a'_t). \tag{21}$$

CACTI instead predicts $\widehat{s}_{t+1}(a'_t) = s_{t+1} + B_\theta(s_t)d_t$. Thus, once the factual event is matched,

$$\widehat{s}_{t+1}(a'_t) - s^\star_{t+1}(a'_t) = B_\theta(s_t)d_t - \tau_t^\star(a'_t). \tag{22}$$

Exact factual reconstruction does not guarantee an accurate counterfactual. The remaining error is the gap between the learned action contrast and the true same-event contrast.

We next separate event mismatch from response mismatch. Define the learned transition

$$\widehat{F}_{\theta,\psi}(s, a; u) = s + \mu_\theta(s, a) + g_\psi(s, u), \tag{23}$$

and, for any candidate event representation $\widetilde{u}$,

$$c_t^F(\widetilde{u}) = \widehat{F}_{\theta,\psi}(s_t, a_t; \widetilde{u}) - s_{t+1}. \tag{24}$$

Proposition 2 (Event–response decomposition) *For any $\widetilde{u}$,*

$$\begin{aligned}\widehat{F}_{\theta,\psi}(s_t, a'_t; \widetilde{u}) - F^\star(s_t, a'_t, u_{t+1}) \\ = c_t^F(\widetilde{u}) + B_\theta(s_t)d_t - \tau_t^\star(a'_t).\end{aligned} \tag{25}$$

The first term measures whether $\widetilde{u}$ still matches the factual event. The second measures whether the learned action contrast matches the true same-event contrast. For the event representation abducted from the factual transition, exact inversion gives $c_t^F(\widehat{u}_{t+1}) = 0$, and Eq. (25) reduces to Eq. (22). The full proof is given in Appendix B.3.

We now bound the response error in Eq. (22). Assume $F^\star(s_t, a, u_{t+1})$ is differentiable in a along the segment from a_t to a'_t, and its action Jacobian is L_t-Lipschitz on that segment. Define

$$\eta_t = \|B_\theta(s_t) - J_A F^\star(s_t, a_t, u_{t+1})\|_{\mathrm{op}}. \tag{26}$$

Here, η_t measures how well the shared response matches the local action response of the factual event.

Proposition 3 (Local counterfactual error) *For a matched factual event,*

$$\left\|\widehat{s}_{t+1}(a'_t) - s^\star_{t+1}(a'_t)\right\|_2 \le \eta_t \|d_t\|_2 + \frac{L_t}{2}\|d_t\|_2^2. \tag{27}$$

The two terms capture different failure modes. The first grows linearly with the mismatch between the shared and local responses. The second is the nonlinear remainder and grows quadratically with the action displacement. Small local mismatch and short interventions therefore favor accurate counterfactual approximation (Appendix B.4).

Together, Sections 4.1 and 4.2 separate two questions. Directional coverage determines whether the target response is identifiable from offline data; local response mismatch and action distance determine how accurately the shared response approximates it.

5 Experiments

We organize the experiments around three questions: Can CACTI predict same-event counterfactual transitions more accurately? Does its prediction depend on information tied to the factual event? Do the resulting transitions help policy learning in low-data regimes? We evaluate MountainCar11, PointMaze, Reacher, and Shadow Hand RotateZ, covering ordered discrete control, goal-conditioned control, nonlinear continuous control, and high-dimensional control.

Same-event oracle. For each factual transition, we restore the simulator state before the action, reuse the disturbance realized in that transition, and change only the action. This gives the same-event counterfactual target for the query action. All methods use the same factual transitions, query actions, and oracle targets. The realized disturbance is used only to construct evaluation targets and is never given to the learned models (Appendix C.1).

Baselines. We compare against direct predictive, probabilistic generative, and explicit-abduction models: Base-D, Base-M (Bishop, 1994), Gaussian dynamics, Conditional Flow (Dinh et al., 2017), BiCoGAN (Lu et al., 2020), and the Tsirtsis location–scale SCM (Tsirtsis & Rodriguez, 2023). These methods differ mainly in how they carry information from the factual event to the query action. Gaussian dynamics and Tsirtsis reuse an abducted factual residual. Conditional Flow inverts under the factual state–action pair and decodes under the query action. BiCoGAN reuses a latent code inferred from the factual transition. Full baseline details are given in Appendix D.2.

Metric. Our primary metric is the normalized mean-squared error between the predicted counterfactual next state and the same-event oracle target:

$$\text{CF-MSE}(\mathcal{Q}) = \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}}\frac{1}{d_s}\left\|D_{\mathrm{eval}}^{-1}\left(\widehat{s}_q^{+} - s_{q,\mathrm{oracle}}^{+}\right)\right\|_2^2. \tag{28}$$

All methods use the same query set and evaluation scale within each environment (Appendix C.2). We report results over all nonfactual queries and over a low-direct-support subset. MountainCar11 measures direct support by empirical action frequency, while the continuous-control tasks use action density (Appendix C.3). These bins stratify empirical query support and are distinct from the directional identifiability condition in Section 4.1. Results are averaged over model seeds $0, 1, 2$; additional support-stratified and state-wise results are reported in Appendix E.

Method	MC11	PointMaze	Reacher	RotateZ
Base-D	1.40×10^{-3} 1.67×10^{-3}	4.31×10^{-4} 6.40×10^{-4}	1.14×10^{-4} 2.35×10^{-4}	0.598 0.767
Gaussian	1.13×10^{-4} 1.88×10^{-4}	3.68×10^{-4} 8.45×10^{-4}	6.48×10^{-5} 1.53×10^{-4}	0.458 0.835
Base-M	1.63×10^{-3} 1.79×10^{-3}	4.64×10^{-4} 6.12×10^{-4}	1.20×10^{-4} 2.73×10^{-4}	0.487 0.742
Cond. Flow	3.04×10^{-4} 7.29×10^{-4}	3.24×10^{-3} 1.51×10^{-3}	1.15×10^{-3} 1.52×10^{-3}	0.498 0.956
BiCoGAN	1.84×10^{-2} 3.32×10^{-2}	1.07×10^{-2} 2.32×10^{-2}	9.04×10^{-2} 1.78×10^{-1}	1.302 2.405
Tsirtsis	1.20×10^{-4} 2.45×10^{-4}	2.05×10^{-4} 3.01×10^{-4}	1.86×10^{-3} 4.35×10^{-3}	1.709 3.254
CACTI	$\mathbf{4.73\times10^{-6}}$ $\mathbf{8.25\times10^{-6}}$	$\mathbf{5.12\times10^{-6}}$ $\mathbf{1.21\times10^{-5}}$	$\mathbf{4.07\times10^{-5}}$ $\mathbf{7.92\times10^{-5}}$	**0.361** **0.662**

Table 1: Same-event counterfactual fidelity. Each cell reports normalized CF-MSE on all nonfactual queries (top) and low-direct-support queries (bottom); lower is better.

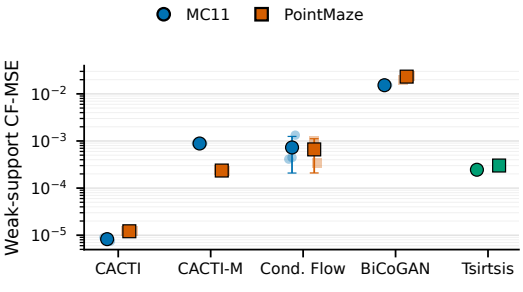

Figure 2: Structural ablation on MountainCar11 and PointMaze. We report low-direct-support CF-MSE; lower is better.

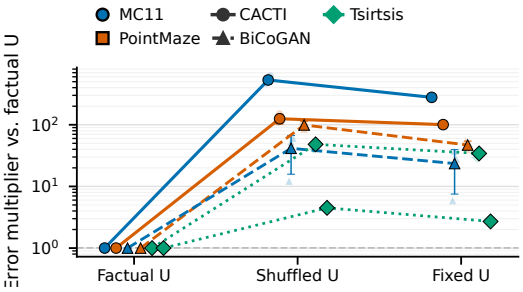

Figure 3: Inferred-U diagnostic. Larger multipliers under shuffled or fixed U indicate stronger degradation from disrupting event-specific correspondence.

5.1 Same-Event Counterfactual Accuracy

Table 1 shows that CACTI achieves the lowest mean same-event counterfactual error in all four environments, both over all nonfactual queries and over the low-direct-support subsets. On MountainCar11 and PointMaze, its all-query error is about $1/24$ and $1/40$ of the strongest non-CACTI baseline, respectively. The gains remain visible on the more challenging Reacher and RotateZ tasks, where CACTI reduces error by about 37% and 21%. The advantage also persists as direct action support decreases. On MountainCar11, for example, errors rise toward less-supported actions, yet CACTI remains below the competing methods across the full action grid.

These differences are not explained by factual reconstruction alone. On MountainCar11, CACTI, Gaussian dynamics, Conditional Flow, and Tsirtsis all reconstruct factual transitions to near numerical precision, yet their counterfactual errors differ substantially. Factual reconstruction is not counterfactual fidelity. A model can explain what happened accurately and still predict the wrong outcome for the same event under a different action.

5.2 Factual Event Matching Matters

CACTI uses both a shared action response and information specific to the factual event. To test whether that event-specific information matters, we compare CACTI with CACTI-M (Figure 2). CACTI-M keeps the mean response $m_\theta(s) + B_\theta(s)a$ but removes the event-specific residual at counterfactual prediction time. On the low-direct-support subsets of MountainCar11 and PointMaze, CACTI-M reaches mean errors of 8.84×10^{-4} and 2.36×10^{-4}, while full CACTI reduces them to 8.25×10^{-6} and 1.21×10^{-5}, respectively. The shared response determines how the outcome changes with the action; same-event prediction also needs information tied to the factual event.

We next keep the trained model, factual transition, and query action fixed, and change only the event representation used at prediction time. Factual U uses the representation inferred from the

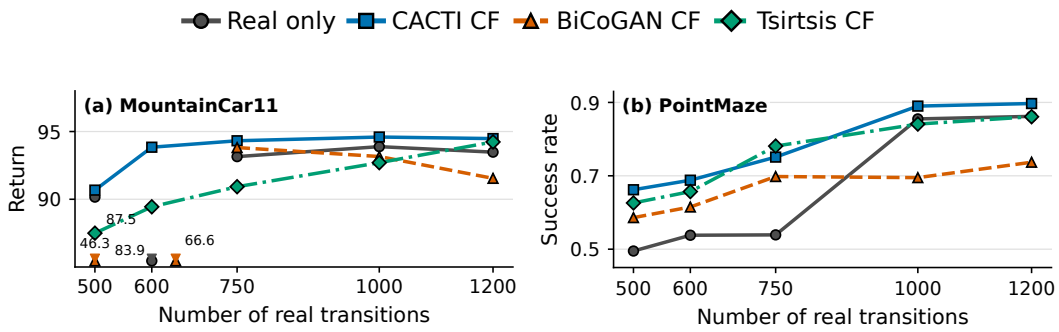

Figure 4: Downstream utility across five real-data budgets. Mean ± standard deviation over policy seeds `0,1,2`; higher is better.

corresponding factual transition. Shuffled U swaps representations across events, preserving their empirical distribution while breaking the factual pairing. Fixed U uses the same latent value for every query. Both perturbations increase CACTI's counterfactual error (Figure 3; Appendix E.3). A plausible event representation is not enough; it must match the factual event. This is exactly the event-matching term in Section 4.2 (Appendix B.5): changing the event representation changes c_t^F while leaving the query action and shared response fixed.

5.3 Downstream Policy Learning under Limited Data

We finally test whether the generated counterfactual transitions help offline policy learning when real data are scarce (Figure 4; Appendix F.1). On MountainCar11 and PointMaze, we use $N_{\text{real}} \in \{500, 600, 750, 1000, 1200\}$ real transitions and fix the synthetic-to-real ratio at 0.5. All methods use the same factual subset, and all augmentation methods use the same query actions. We compare real-only training with augmentation from CACTI, BiCoGAN, and Tsirtsis, and report mean ± standard deviation over three policy seeds.

CACTI gives the strongest average downstream performance in most settings. On MountainCar11, it achieves the highest mean return at all five data budgets. At $N_{\text{real}} = 500$, the mean returns are 90.67 for CACTI, 90.16 for real-only training, 87.50 for Tsirtsis, and 46.25 for BiCoGAN. The gaps narrow as the task approaches its performance ceiling. On PointMaze, CACTI achieves the highest mean success rate in four of five budgets. At $N_{\text{real}} = 1200$, its success rate reaches 0.897, compared with 0.862 for real-only training, 0.861 for Tsirtsis, and 0.737 for BiCoGAN. Across the ten settings, CACTI attains the highest mean primary metric in nine. Under these low-data settings, CACTI-generated counterfactuals provide useful training signal, although the size of the gain depends on the task and data budget.

6 Conclusion

We study same-event counterfactual transition inference in offline reinforcement learning. CACTI combines a shared state-dependent action response with a conditional invertible model of event-specific variation. For one-step queries, the factual transition preserves the event contribution, while the shared response determines the effect of changing the action. We show when the target action response is identifiable from offline data and how local response mismatch and action distance control counterfactual error. Across four control tasks, CACTI achieves the lowest mean same-event counterfactual error on both all queries and low-direct-support subsets. Event-matching diagnostics and low-data policy experiments further support the usefulness of the generated transitions. The method relies on sequential exogeneity, shared action response, and an action-independent exogenous term, and our study is limited to one-step counterfactuals.

AI USE STATEMENT

We used generative AI tools to assist with manuscript organization and editing, translation, and literature search. All mathematical claims, proofs, citations, experimental procedures, and reported results were reviewed and verified by the authors. We take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

We provide the model definition, training objective, and same-event query procedure in the main paper. Appendix A gives the exact likelihood, training schedule, and normalization details, and Appendix D.1 describes the architectures. Appendix B contains complete proofs of the theoretical results. Appendices C and D specify the environments, same-event oracle, evaluation metric, baseline query rules, implementation details, random seeds, and checkpoint selection. Appendix E reports additional support-stratified, state-wise, and event-matching diagnostics, and Appendix F documents the downstream policy-learning protocol. Code and experiment configurations are included in the supplementary material.

REFERENCES

Christopher Bishop. Mixture density networks. Technical Report NCRG/94/004, January 1994. URL https://www.microsoft.com/en-us/research/publication/mixture-density-networks/.

Yikang Chen and Dehui Du. Exogenous isomorphism for counterfactual identifiability. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 8022–8064. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/chen25p.html.

Minjae Cho and Chuangchuang Sun. Out of distribution adaptation in offline rl via causal normalizing flows. *Mathematics*, 13(23), 2025. ISSN 2227-7390. doi: 10.3390/math13233835. URL https://www.mdpi.com/2227-7390/13/23/3835.

Fabio De Sousa Ribeiro, Tian Xia, Miguel Monteiro, Nick Pawlowski, and Ben Glocker. High fidelity image counterfactuals with probabilistic causal models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7390–7425. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/de-sousa-ribeiro23a.html.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp, 2017. URL https://arxiv.org/abs/1605.08803.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning, 2021. URL https://arxiv.org/abs/2004.07219.

Scott Fujimoto and Shixiang (Shane) Gu. A minimalist approach to offline reinforcement learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 20132–20145. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/a8166da05c5a094f7dc03724b41886e5-Paper.pdf.

Alexander Immer, Christoph Schultheiss, Julia E Vogt, Bernhard Schölkopf, Peter Bühlmann, and Alexander Marx. On the identifiability and estimation of causal location-scale noise models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 14316–14332. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/immer23a.html.

Milad Kazemi, Jessica Lally, Ekaterina Tishchenko, Hana Chockler, and Nicola Paoletti. Counterfactual influence in markov decision processes. In Biwei Huang and Mathias Drton (eds.), *Proceedings of the Fourth Conference on Causal Learning and Reasoning*, volume 275 of *Proceedings of Machine Learning Research*, pp. 792–817. PMLR, 07–09 May 2025. URL https://proceedings.mlr.press/v275/kazemi25a.html.

Ilyes Khemakhem, Ricardo Monti, Robert Leech, and Aapo Hyvarinen. Causal autoregressive flows. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3520–3528. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/khemakhem21a.html.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1179–1191. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/0d2b2061826a5df3221116a5085a6052-Paper.pdf.

Jessica Lally, Milad Kazemi, and Nicola Paoletti. Robust counterfactual inference in markov decision processes, 2026. URL https://arxiv.org/abs/2502.13731.

Jaewoo Lee, Sujin Yun, Taeyoung Yun, and Jinkyoo Park. Gta: Generative trajectory augmentation with guidance for offline reinforcement learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 56766–56801. Curran Associates, Inc., 2024. doi: 10.52202/079017-1808. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/67ea314d1df751bbf99ab664ae3049a5-Paper-Conference.pdf.

Chaochao Lu, Biwei Huang, Ke Wang, José Miguel Hernández-Lobato, Kun Zhang, and Bernhard Schölkopf. Sample-efficient reinforcement learning via counterfactual-based data augmentation, 2020. URL https://arxiv.org/abs/2012.09092.

Arash Nasr-Esfahany and Emre Kiciman. Counterfactual (non-)identifiability of learned structural causal models, 2023. URL https://arxiv.org/abs/2301.09031.

Arash Nasr-Esfahany, Mohammad Alizadeh, and Devavrat Shah. Counterfactual identifiability of bijective causal models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25733–25754. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/nasr-esfahany23a.html.

Kimia Noorbakhsh and Manuel Rodriguez. Counterfactual temporal point processes. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24810–24823. Curran Associates, Inc., 2022. doi: 10.52202/068431-1799. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d3faa41886997cfc2128b930077fa49-Paper-Conference.pdf.

Michael Oberst and David Sontag. Counterfactual off-policy evaluation with Gumbel-max structural causal models. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4881–4890. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/oberst19a.html.

Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 857–869. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/0987b8b338d6c90bbedd8631bc499221-Paper.pdf.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference - Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning Series. The MIT Press,

Cambridge, MA, USA, 2017. URL https://library.oapen.org/bitstream/id/056a11be-ce3a-44b9-8987-a6c68fce8d9b/11283.pdf.

Silviu Pitis, Elliot Creager, and Animesh Garg. Counterfactual data augmentation using locally factored dynamics. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3976–3990. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/294e09f267683c7ddc6cc5134a7e68a8-Paper.pdf.

Silviu Pitis, Elliot Creager, Ajay Mandlekar, and Animesh Garg. Mocoda: Model-based counterfactual data augmentation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 18143–18156. Curran Associates, Inc., 2022. doi: 10.52202/068431-1319. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/7314e20a73542bbfff25030d1185ce88-Paper-Conference.pdf.

Jonathan Richens, Rory Beard, and Daniel H. Thompson. Counterfactual harm. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 36350–36365. Curran Associates, Inc., 2022. doi: 10.52202/068431-2634. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/ebcf1bff7b2fe6dcc3fbe666faaa50f1-Paper-Conference.pdf.

Stratis Tsirtsis and Manuel Rodriguez. Finding counterfactually optimal action sequences in continuous state spaces. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 3220–3247. Curran Associates, Inc., 2023. doi: 10.52202/075280-0143. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/09ae6beae5f1ff38f05c05979097ea0f-Paper-Conference.pdf.

Stratis Tsirtsis, Abir De, and Manuel Rodriguez. Counterfactual explanations in sequential decision making under uncertainty. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 30127–30139. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/fd0a5a5e367a0955d81278062ef37429-Paper.pdf.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 14129–14142. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/a322852ce0df73e204b7e67cbbef0d0a-Paper.pdf.

Tianhe Yu, Aviral Kumar, Rafael Rafailov, Aravind Rajeswaran, Sergey Levine, and Chelsea Finn. Combo: Conservative offline model-based policy optimization. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 28954–28967. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/f29a179746902e331572c483c45e5086-Paper.pdf.

Wei Zhang, Azlaan Mustafa SAMAD, Yuan Xue, and Daniel Kudenko. Causality-based data-augmentation for offline reinforcement learning. *IEEE Access*, 14:76639–76648, 2026. doi: 10.1109/ACCESS.2026.3694183.

SUPPLEMENTARY OVERVIEW

Appendix A gives additional likelihood and training details, with the full architecture in Appendix D.1. Appendix B presents cross-action ambiguity, action-direction identifiability, counterfactual error, and supporting event-consistency identities. Appendices C and D document evaluation and baselines. Appendix E separates direct-support results, semantic diagnostics, event matching, and exploratory variants. Appendix F gives the downstream protocol and scope.

A ADDITIONAL MODEL AND TRAINING DETAILS

We use the transition notation from Sections 3.1–3.3.

A.1 EXACT CONDITIONAL LIKELIHOOD

The implementation models state increments rather than raw next states. Let $\mathcal{N}_\Delta$ denote the invertible increment normalizer used during training and write

$$x_{t+1} = \mathcal{N}_\Delta(\Delta s_{t+1}). \tag{S1}$$

Let $\bar{\mu}_\theta(s, a)$ denote the systematic prediction represented in the same normalized coordinates. The flow residual and its inverse code are

$$z^\theta_{t+1} = x_{t+1} - \bar{\mu}_\theta(s_t, a_t), \qquad u_{t+1} = g^{-1}_{\psi,s_t}(z^\theta_{t+1}). \tag{S2}$$

For reference density p_U, change of variables (Dinh et al., 2017) gives

$$\log p_{\theta,\psi}(x_{t+1} \mid s_t, a_t) = \log p_U(u_{t+1}) \tag{S3}$$

$$+ \log \left| \det J_{g^{-1}_{\psi,s_t}}(z^\theta_{t+1}) \right|. \tag{S4}$$

The corresponding raw-increment density is

$$\log p_{\theta,\psi}(\Delta s_{t+1} \mid s_t, a_t) = \log p_{\theta,\psi}(x_{t+1} \mid s_t, a_t) \tag{S5}$$

$$+ \log \left| \det J_{\mathcal{N}_\Delta}(\Delta s_{t+1}) \right|. \tag{S6}$$

For an affine diagonal normalizer, the final term is constant with respect to the model parameters and may be omitted during optimization, but it remains part of the exact raw-space density. The reference density p_U fixes a numerical coordinate system for the learned exogenous representation; it is not a claim that the physical disturbance itself follows that reference distribution.

Probabilistic causal generative models also support image counterfactual inference (De Sousa Ribeiro et al., 2023).

A.2 TRAINING OBJECTIVE AND SCHEDULE

The factual negative log likelihood is

$$\mathcal{L}_{\text{fact}} = -\mathbb{E}_{(s_t,a_t,s_{t+1})\sim\mathcal{D}} \left[\log p_{\theta,\psi}(\Delta s_{t+1} \mid s_t, a_t) \right]. \tag{S7}$$

We supplement the likelihood with a squared-error objective on the response branch,

$$\mathcal{L}_{\text{resp}} = \mathbb{E}_{\mathcal{D}} \left[\| \Delta s_{t+1} - m_\theta(s_t) - B_\theta(s_t) a_t \|_2^2 \right], \tag{S8}$$

with response regularizer

$$\mathcal{R}_B = \mathbb{E}_{\mathcal{D}} \left[\| B_\theta(s_t) \|_F^2 \right]. \tag{S9}$$

The complete objective is

$$\mathcal{L}_{\text{CACTI}} = \mathcal{L}_{\text{fact}} + \lambda_{\text{resp}} \mathcal{L}_{\text{resp}} + \lambda_B \mathcal{R}_B. \tag{S10}$$

The likelihood also depends on m_θ and B_θ through μ_θ. We first stabilize m_θ and B_θ with the response and regularization terms, and then optimize m_θ, B_θ, and g_ψ jointly with Eq. (S10). We use AdamW and select checkpoints by factual validation negative log likelihood.

A.3 Architecture and Normalization

The response prediction and flow residual use the same normalized increment coordinates defined in Appendix A.1. The raw-increment density includes the normalizer Jacobian in Eq. (S6). Appendix D.1 gives the full CACTI implementation, including the environment-specific network sizes, flow blocks, latent dimensions, and variant definitions. CACTI-S is an exploratory structural-mask variant; its results are collected in Appendix E.4.

B Formal Results and Proofs

B.1 Cross-Action Ambiguity

Fix a state $s_t = s$, and let Y_{t+1} denote one scalar component of the next state, such as the vehicle's lateral position. Suppose the unobserved exogenous variable U_{t+1} has two equally likely realizations, u_1 and u_2. The following two SCMs are consistent with the same observed distributions:

	u_1	u_2
right	3	1
$\mathcal{M}_1$: left	-1	-3
$\mathcal{M}_2$: left	-3	-1

The entries are possible values of Y_{t+1}. Both models predict 3 or 1 under the right action and -1 or -3 under the left action, each with probability $1/2$. However, for a factual right action with $Y_{t+1} = 3$, $\mathcal{M}_1$ predicts the counterfactual left-turn outcome -1, whereas $\mathcal{M}_2$ predicts -3.

The data determine the rows of the table, the possible outcomes under each action, but not its columns, which outcomes belong to the same factual event.

Proposition S1 (Action-conditional fit is insufficient) *Fix a state s. Even when the distribution of $G(s, a, U)$ is identical for every action a, the same-event coupling across actions need not be determined.*

Proof. Let $U \sim P_U$ and, for each action a, let T_a be a bijection that preserves P_U, so $T_a(U) \overset{d}{=} U$. Define

$$\widetilde{G}(s, a, u) = G(s, a, T_a(u)). \tag{S11}$$

For every fixed action,

$$\widetilde{G}(s, a, U) = G(s, a, T_a(U)) \overset{d}{=} G(s, a, U), \tag{S12}$$

so the two SCMs have identical action-conditional outcome distributions.

Now suppose the factual action is a and the factual realization under G is u. The transformed SCM explains the same factual outcome using $\widetilde{u} = T_a^{-1}(u)$. Under query action a', its same-event prediction is

$$\widetilde{G}(s, a', \widetilde{u}) = G\big(s, a', T_{a'} T_a^{-1}(u)\big), \tag{S13}$$

whereas the original SCM predicts $G(s, a', u)$. Whenever $T_{a'} T_a^{-1}$ is not the identity on a set of positive probability and $G(s, a', \cdot)$ distinguishes the affected values, the same-event counterfactuals differ despite identical observational distributions. □

This general construction formalizes the ambiguity in same-event coupling across actions (Nasr-Esfahany & Kiciman, 2023). Sharing one exogenous map across actions removes this specific action-dependent relabeling degree of freedom.

Related work bounds counterfactual transition probabilities across compatible causal models in MDPs (Lally et al., 2026).

B.2 Proof of Theorem 1: Action-Direction Identifiability

At a fixed state s, suppose the shared-response model is exact locally:

$$\Delta S_{t+1} = m^\star(S_t) + B^\star(S_t) A_t + g^\star(S_t, U_{t+1}), \tag{S14}$$

and assume sequential exogeneity $U_{t+1} \perp A_t \mid S_t$. Define

$$\Sigma_A(s) = \mathrm{Cov}(A_t \mid S_t = s), \tag{S15}$$

$$C_{\Delta A}(s) = \mathrm{Cov}(\Delta S_{t+1}, A_t \mid S_t = s). \tag{S16}$$

Theorem S1 (Action-direction identifiability) *Assume the relevant conditional second moments exist. Within the state-local model in Eq.* (S14)*, and without additional cross-state restrictions, the query effect $B^\star(s)d$ is uniquely determined by the observational distribution if and only if*

$$d \in \mathrm{Range}(\Sigma_A(s)). \tag{S17}$$

When this condition holds,

$$B^\star(s)d = C_{\Delta A}(s)\Sigma_A(s)^\dagger d. \tag{S18}$$

Proof: sufficiency. Sequential exogeneity implies

$$\mathrm{Cov}(g^\star(s, U_{t+1}), A_t \mid S_t = s) = 0, \tag{S19}$$

so

$$C_{\Delta A}(s) = B^\star(s)\Sigma_A(s). \tag{S20}$$

If $d \in \mathrm{Range}(\Sigma_A(s))$, then $\Sigma_A(s)\Sigma_A(s)^\dagger d = d$. Multiplying Eq. (S20) by $\Sigma_A(s)^\dagger d$ gives Eq. (S18). Hence every response matrix consistent with the observed conditional moments agrees on the query effect.

Proof: necessity. Fix s and write $\bar{a} = \mathbb{E}[A_t \mid S_t = s]$ and $\Sigma = \Sigma_A(s)$. Suppose $d \notin \mathrm{Range}(\Sigma)$. Symmetry and positive semidefiniteness give the orthogonal decomposition $d = d_R + d_N$, where $d_R \in \mathrm{Range}(\Sigma)$ and $0 \neq d_N \in \mathrm{Null}(\Sigma)$. Let d_s and d_a be the state and action dimensions. For any nonzero $v \in \mathbb{R}^{d_s}$, define

$$H = \frac{v d_N^\top}{\|d_N\|_2^2} \in \mathbb{R}^{d_s \times d_a}. \tag{S21}$$

Then $Hd = v \neq 0$. Since

$$\mathrm{Var}(d_N^\top A_t \mid S_t = s) = d_N^\top \Sigma d_N = 0,$$

we have $d_N^\top(A_t - \bar{a}) = 0$ and hence $H(A_t - \bar{a}) = 0$ almost surely conditional on $S_t = s$.

Define an alternative SCM at s by

$$\widetilde{B}(s) = B^\star(s) + H, \qquad \widetilde{m}(s) = m^\star(s) - H\bar{a}, \qquad \widetilde{g}(s, u) = g^\star(s, u),$$

and keep the conditional law of (A_t, U_{t+1}) given $S_t = s$ unchanged. Under this law,

$$\begin{aligned}
&\widetilde{m}(s) + \widetilde{B}(s)A_t + \widetilde{g}(s, U_{t+1}) \\
&\quad = m^\star(s) + B^\star(s)A_t + g^\star(s, U_{t+1}) + H(A_t - \bar{a}) \\
&\quad = m^\star(s) + B^\star(s)A_t + g^\star(s, U_{t+1})
\end{aligned}$$

almost surely conditional on $S_t = s$. Thus the two SCMs induce the same full conditional observational distribution of $(A_t, \Delta S_{t+1}) \mid S_t = s$, while

$$\widetilde{B}(s)d - B^\star(s)d = Hd = v \neq 0.$$

The alternative remains in the same state-local shared-response model class. The unchanged conditional law preserves sequential exogeneity, and the unchanged exogenous map preserves any assumed invertibility of $g^\star$. Hence the query effect is not unique within this class. The construction is local to s; it makes no non-identifiability claim under additional cross-state parameter sharing or other structural restrictions. □

Remark S1 *If $\Sigma_A(s) \succ 0$, every local direction is covered. If it is singular, only directions in its range are identified. Directional coverage does not guarantee finite-sample stability or accuracy for a large action displacement; those are separate estimation and local-approximation questions.*

B.3 Proof of Proposition 2: Event–Response Decomposition

Use the learned transition, factual discrepancy, and true same-event contrast defined in Eqs. (23), (24), and (20), with $d_t = a'_t - a_t$. We write $\widehat{F}$ for $\widehat{F}_{\theta,\psi}$ in the proof.

Proof. Add and subtract the learned factual prediction:

$$\widehat{F}(s_t, a'_t; \widetilde{u}) - F^\star(s_t, a'_t, u_{t+1}) \tag{S22}$$

$$= \left[\widehat{F}(s_t, a_t; \widetilde{u}) - F^\star(s_t, a_t, u_{t+1})\right] \tag{S23}$$

$$+ \left[\widehat{F}(s_t, a'_t; \widetilde{u}) - \widehat{F}(s_t, a_t; \widetilde{u})\right] - \tau_t^\star(a'_t). \tag{S24}$$

Because $F^\star(s_t, a_t, u_{t+1}) = s_{t+1}$, the first bracket is $c_t^{\mathrm{F}}(\widetilde{u})$. Because the learned exogenous contribution is action independent, the second bracket is $B_\theta(s_t)(a'_t - a_t)$. □

B.4 Proof of Proposition 3: Local Counterfactual Error

Proposition S2 (Local action-space error bound) *Assume $a \mapsto F^\star(s_t, a, u_{t+1})$ is differentiable on the line segment from a_t to a'_t and its action Jacobian is L_t-Lipschitz there. Let η_t be the response mismatch defined in Eq.* (26)*. Then, for any candidate event representation $\widetilde{u}$,*

$$\left\|\widehat{F}(s_t, a'_t; \widetilde{u}) - F^\star(s_t, a'_t, u_{t+1})\right\|_2 \tag{S25}$$

$$\leq \left\|c_t^{\mathrm{F}}(\widetilde{u})\right\|_2 + \eta_t \left\|d_t\right\|_2 + \frac{L_t}{2}\left\|d_t\right\|_2^2. \tag{S26}$$

Proof. By Proposition 2 and the triangle inequality, it is enough to bound $\|B_\theta(s_t)d_t - \tau_t^\star(a'_t)\|_2$. Taylor's theorem with Lipschitz Jacobian gives

$$\tau_t^\star(a'_t) = J_A F^\star(s_t, a_t, u_{t+1})d_t + r_t, \tag{S27}$$

$$\|r_t\|_2 \leq \frac{L_t}{2}\left\|d_t\right\|_2^2. \tag{S28}$$

Therefore

$$\|B_\theta(s_t)d_t - \tau_t^\star(a'_t)\|_2 \leq \eta_t \left\|d_t\right\|_2 + \left\|r_t\right\|_2, \tag{S29}$$

which yields Eq. (S26). □

For the matched event representation $\widetilde{u} = \widehat{u}_{t+1}$, exact factual consistency gives $c_t^{\mathrm{F}}(\widehat{u}_{t+1}) = 0$ and removes the first term. This gives Eq. (27) and proves Proposition 3. The mismatch η_t may include structural approximation error: B_θ depends only on state, while the true local action response may depend on the action and factual event. The remaining term bounds the nonlinear remainder as action distance grows.

B.5 Factual Consistency and Event-Perturbation Identities

Corollary S1 (Uniqueness of the factual event representation) *If g_{ψ,s_t} is injective and $\widehat{u}_{t+1}$ is defined by Eq.* (11)*, then*

$$c_t^{\mathrm{F}}(\widetilde{u}) = 0 \quad \Longleftrightarrow \quad \widetilde{u} = \widehat{u}_{t+1}. \tag{S30}$$

Proof. Exact inversion gives

$$c_t^{\mathrm{F}}(\widetilde{u}) = g_\psi(s_t, \widetilde{u}) - g_\psi(s_t, \widehat{u}_{t+1}). \tag{S31}$$

The difference is zero if and only if the two inputs are equal by injectivity. □

Corollary S2 (Event-perturbation identity) *For every query action a'_t and candidate event representation $\widetilde{u}$,*

$$\widehat{F}(s_t, a'_t; \widetilde{u}) - \widehat{F}(s_t, a'_t; \widehat{u}_{t+1}) = c_t^{\mathrm{F}}(\widetilde{u}). \tag{S32}$$

Proof. The shared state and action terms cancel, leaving $g_\psi(s_t, \widetilde{u}) - g_\psi(s_t, \widehat{u}_{t+1})$, which equals the factual-consistency discrepancy above. □

Equation (S32) explains the diagnostic intervention: shuffled or fixed event representations alter factual-event correspondence while keeping the factual state, query action, and learned shared response unchanged. It does not imply pointwise degradation for every query, because the perturbation

can occasionally cancel another model error; the experiment tests systematic average degradation. These are model-space factual-consistency statements. They do not imply numerical recovery of the physical disturbance.

Trajectory-level work also studies whether factual observations continue to influence counterfactual paths (Kazemi et al., 2025).

C Environments and Same-Event Evaluation

C.1 Same-Event Simulator Oracle

The same-event target changes only the queried action. The oracle disturbance is used to construct evaluation targets and is never provided to the learned models.

Algorithm 1 Same-event simulator oracle for one factual event

Require: Simulator snapshot before the factual step; factual action a_t; query set $\mathcal{A}'_t$
1: Save the physical state, simulator internals, and random-number-generator state.
2: Execute the factual action once and record s_{t+1} and the realized step disturbance, either explicitly or through the restorable RNG state.
3: **for** each $a'_t \in \mathcal{A}'_t$ **do**
4: Restore the exact pre-step simulator and RNG state.
5: Reuse the factual step disturbance and replace only a_t by a'_t.
6: Execute one step and record the oracle target $s^{\text{oracle}}_{t+1}(a'_t)$.
7: **end for**
8: Restore the simulator so that one query cannot alter another query.

All methods are evaluated on the same factual test transitions, query actions, and oracle targets. The learned model sees only (s_t, a_t, s_{t+1}) and must infer its own event representation $\widehat{U}$.

C.2 Counterfactual Metric and Normalization

Let D_{eval} be the environment-specific diagonal scale used by the evaluator and let $\mathcal{Q}$ contain non-factual query pairs. Here, d_s is the state dimension, $\widehat{s}^+_q$ is the predicted counterfactual next state for query q, and $s^+_{q,\text{oracle}}$ is its same-event oracle target. The primary metric is

$$\text{CF-MSE}(\mathcal{Q}) = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{1}{d_s} \left\| D^{-1}_{\text{eval}} (\widehat{s}^+_q - s^+_{q,\text{oracle}}) \right\|^2_2 . \tag{S33}$$

The main paper reports both all-query and low-direct-support averages. Support-stratified tables use the same evaluator and differ only in the query subset. Every method uses the same query set and environment-specific evaluation scale.

C.3 Environments and Direct-Support Bins

MountainCar11 provides ordered discrete control, PointMaze provides goal-conditioned continuous control, Reacher provides nonlinear continuous control, and Shadow Hand RotateZ provides high-dimensional dexterous control. Support is measured by empirical action frequency in MountainCar11 and by the common action-density evaluator in the continuous environments.

MountainCar11. The ordered action grid is

$$\{-1.0, -0.8, -0.6, -0.4, -0.2, 0, 0.2, 0.4, 0.6, 0.8, 1.0\}. \tag{S34}$$

The action-wise evaluation labels $\{-1.0, 1.0\}$ as low direct support, $\{-0.8, 0.6, 0.8\}$ as medium, and the remaining grid values as strong.

PointMaze. The flat state used in the blockwise evaluation is

$$s = [x, y, v_x, v_y, g_x, g_y]. \tag{S35}$$

The desired-goal coordinates (g_x, g_y) are invariant under one environment step. The CACTI-S diagnostic enforces this relation with a hard structural prior. For continuous tasks, the tables report High, Medium, and Low action-density regions; Low denotes the low-direct-support subset used in the main paper.

Reacher and RotateZ. The same one-step same-event oracle is used for both environments. For Reacher, support is the state-conditional behavior-policy log density. High, Medium, and Low are the top 30%, middle 40%, and bottom 30% of queries, respectively; each model seed is evaluated on $640{,}000$ queries. Reacher support-stratified and state-block results appear in Appendix E; the RotateZ all-query and low-direct-support results remain in the main paper.

These bins stratify empirical direct query support. They are distinct from the directional identifiability condition in Section 4.1.

C.4 Evaluation Fairness

The main-paper comparison uses identical factual test transitions and alternative-action queries across methods and averages over model seeds $0, 1, 2$. The same oracle targets and environment-specific evaluation normalization are used for every method.

D Implementation and Baselines

D.1 CACTI Architecture

CACTI. Throughout the paper, CACTI denotes the no-mask model. The state-only component m_θ and the state-dependent action response B_θ are implemented as separate multilayer perceptrons with three hidden layers and SiLU activations. For MountainCar11, B_θ outputs a state-dimensional response vector that is multiplied by the scalar action. For the continuous-control environments, B_θ outputs a $d_s \times d_a$ response matrix that is multiplied by the action vector. The numerical action is used directly, without a learned action embedding.

The event-specific component g_ψ is implemented by a state-conditioned RealNVP (Dinh et al., 2017). Each coupling block uses a two-hidden-layer SiLU conditioner. The flow is conditioned on the current state but not on the action; action dependence is represented exclusively by $B_\theta(s)a$. The latent dimension equals the state-increment dimension.

Table S1: Environment-specific CACTI architectures. An entry $L \times H$ denotes L hidden layers of width H. The RealNVP latent dimension equals d_s.

Environment	d_s/d_a	m_θ MLP	B_θ MLP	Coupling MLP	Flow blocks
MountainCar11	$2/1$	3×256	3×256	2×256	6
PointMaze	$6/2$	3×256	3×256	2×256	8
Reacher	$10/2$	3×256	3×256	2×256	8
RotateZ	$68/20$	3×384	3×384	2×384	8

Variants. CACTI-S uses the same backbone and applies a predefined structural mask to the action-response output $B_\theta(s)$. CACTI-M retains the mean-response branch $m_\theta(s) + B_\theta(s)a$ but omits the event-specific residual contribution during counterfactual prediction. The training objective, warm-up schedule, optimizer, and checkpoint selection are given in Appendix A.2.

D.2 Baseline Definitions and Same-Event Query Rules

Each baseline is learned from factual transitions. We describe its fitted transition model, the information taken from the factual event, and its query rule.

Base-D. This deterministic dynamics baseline learns the next-state increment from (s, a). It infers no event-specific representation. At query time it evaluates the same predictor at (s_t, a'_t); the factual next state does not provide a residual to reuse.

Gaussian dynamics. This conditional diagonal-Gaussian transition model learns a conditional mean and scale from factual transitions (Pawlowski et al., 2020; Immer et al., 2023). It infers a standardized residual from the factual transition. Under the query action, it evaluates the conditional mean and scale at (s_t, a'_t) and reuses that residual.

Base-M. This mixture-density predictive baseline learns the conditional transition distribution (Bishop, 1994). Its deterministic prediction uses the mixture expectation at the query state–action pair, without abducting an event representation from the factual next state. Stochastic sampling is treated only as a diagnostic.

Conditional Flow. This action-conditioned RealNVP learns the factual conditional transition distribution (Dinh et al., 2017). It inverts under the factual (s_t, a_t) and decodes the resulting code under (s_t, a'_t). Unlike CACTI, the action enters the flow conditioner directly.

BiCoGAN. The Lu-style explicit-abduction baseline jointly learns an encoder, generator, and discriminator (Lu et al., 2020). It infers a latent code from the factual transition and reuses that code after changing the action, without CACTI's shared control-response decomposition.

Tsirtsis location–scale SCM. The transition component (Tsirtsis & Rodriguez, 2023) is implemented as

$$Y = h_\theta(s, a) + \phi_\theta(s, a) \odot U, \tag{S36}$$

where h_θ and ϕ_θ are the learned location and scale functions. Factual abduction gives

$$\widehat{U} = \frac{Y - h_\theta(s, a)}{\phi_\theta(s, a)}. \tag{S37}$$

The query evaluates the location and scale at (s_t, a'_t) and reuses the abducted $\widehat{U}$. The comparison uses this transition SCM, not the complete counterfactually optimal action-sequence search procedure.

D.3 Reproducibility and Compute

All experiments were run on one NVIDIA GeForce RTX 3090 GPU. Counterfactual-fidelity results use model seeds $0, 1, 2$. The low-data policy experiments in the main paper use policy seeds $0, 1, 2$. Checkpoints are selected by factual validation objectives rather than oracle counterfactual error.

E Additional Counterfactual Results

The main paper reports the four-environment all-query and low-direct-support comparison. This appendix provides the additional support, action, state-block, semantic, and event-matching diagnostics.

E.1 Support- and Action-Stratified Results

Tables S2, S3, S4, and S5 report empirical direct-support and action-stratified results. Low denotes low direct support, as defined in Appendix C.3.

MountainCar11. All methods become less accurate as direct support decreases, but CACTI remains the most accurate in every reported tier. CACTI-M changes little across tiers because it predicts an average response; this stability should not be confused with low event-level error. The action-wise trend is smooth rather than being driven by one isolated query: errors increase toward the sparsely observed extremes, while CACTI remains below the alternatives throughout the action grid.

Table S2: MountainCar11 normalized CF-MSE. All is the aggregate over nonfactual queries; support-tier entries are mean ± sample standard deviation over model seeds $0, 1, 2$ where available. Lower is better.

Method	All	Strong	Medium	Low
CACTI	$\mathbf{4.73\times10^{-6}}$	$\mathbf{(2.55\pm0.18)\times10^{-6}}$	$\mathbf{(6.46\pm0.63)\times10^{-6}}$	$\mathbf{(8.25\pm0.74)\times10^{-6}}$
CACTI-S	5.10×10^{-5}	$(2.44\pm3.60)\times10^{-5}$	$(7.32\pm11.3)\times10^{-5}$	$(9.24\pm14.1)\times10^{-5}$
Gaussian	1.13×10^{-4}	$(5.44\pm1.03)\times10^{-5}$	$(1.72\pm0.46)\times10^{-4}$	$(1.88\pm0.39)\times10^{-4}$
Conditional Flow	3.04×10^{-4}	$(6.60\pm6.84)\times10^{-5}$	$(4.67\pm4.01)\times10^{-4}$	$(7.29\pm5.20)\times10^{-4}$
CACTI-M	8.70×10^{-4}	$(8.62\pm0.12)\times10^{-4}$	$(8.75\pm0.08)\times10^{-4}$	$(8.84\pm0.05)\times10^{-4}$
BiCoGAN	1.84×10^{-2}	$(1.09\pm1.40)\times10^{-2}$	$(2.23\pm3.05)\times10^{-2}$	$(3.32\pm4.47)\times10^{-2}$
Tsirtsis	1.20×10^{-4}	$(5.67\pm0.55)\times10^{-5}$	$(1.56\pm0.31)\times10^{-4}$	$(2.45\pm0.28)\times10^{-4}$

Table S3: MountainCar11 action-wise normalized oracle CF-MSE, averaged over model seeds $0, 1, 2$. Lower is better.

Action	Support	Pairs	CACTI	CACTI-S	Gaussian	Cond. Flow	BiCoGAN
−1.0	low	19210	4.53×10^{-6}	3.85×10^{-5}	6.00×10^{-5}	4.69×10^{-5}	2.97×10^{-2}
−0.8	medium	18384	3.01×10^{-6}	2.35×10^{-5}	3.98×10^{-5}	2.89×10^{-5}	2.26×10^{-2}
−0.6	strong	17597	2.01×10^{-6}	1.37×10^{-5}	4.13×10^{-5}	2.56×10^{-5}	1.68×10^{-2}
−0.4	strong	17406	1.50×10^{-6}	9.55×10^{-6}	4.01×10^{-5}	2.15×10^{-5}	1.20×10^{-2}
−0.2	strong	17668	1.52×10^{-6}	1.16×10^{-5}	3.10×10^{-5}	1.75×10^{-5}	8.69×10^{-3}
0.0	strong	17980	2.08×10^{-6}	1.99×10^{-5}	3.10×10^{-5}	2.26×10^{-5}	7.48×10^{-3}
0.2	strong	17919	3.23×10^{-6}	3.48×10^{-5}	5.97×10^{-5}	6.76×10^{-5}	8.64×10^{-3}
0.4	strong	17760	4.94×10^{-6}	5.65×10^{-5}	1.23×10^{-4}	2.40×10^{-4}	1.21×10^{-2}
0.6	medium	18070	7.01×10^{-6}	8.29×10^{-5}	2.03×10^{-4}	4.74×10^{-4}	1.80×10^{-2}
0.8	medium	18669	9.33×10^{-6}	1.13×10^{-4}	2.71×10^{-4}	8.93×10^{-4}	2.62×10^{-2}
1.0	low	19337	1.19×10^{-5}	1.46×10^{-4}	3.15×10^{-4}	1.41×10^{-3}	3.67×10^{-2}

PointMaze. CACTI remains the most accurate method in the High, Medium, and Low regions. Conditional Flow shows substantial seed variability, while the explicit-abduction baselines remain well above CACTI under low direct support.

Table S4: PointMaze normalized CF-MSE. All is the aggregate over nonfactual queries; support-region entries are mean ± sample standard deviation over three model seeds where available. Low denotes the low-direct-support subset used in the main paper. Lower is better.

Method	All	High	Medium	Low
CACTI	$\mathbf{5.12\times10^{-6}}$	$\mathbf{(7.85\pm0.01)\times10^{-7}}$	$\mathbf{(3.13\pm0.00)\times10^{-6}}$	$\mathbf{(1.21\pm0.00)\times10^{-5}}$
CACTI-S	5.73×10^{-6}	$(9.43\pm0.43)\times10^{-7}$	$(3.54\pm0.10)\times10^{-6}$	$(1.34\pm0.03)\times10^{-5}$
CACTI-M	2.05×10^{-4}	$(1.63\pm0.08)\times10^{-4}$	$(2.13\pm0.11)\times10^{-4}$	$(2.36\pm0.09)\times10^{-4}$
Gaussian	3.68×10^{-4}	$(7.99\pm1.75)\times10^{-5}$	$(2.25\pm0.65)\times10^{-4}$	$(8.45\pm4.25)\times10^{-4}$
Conditional Flow	3.24×10^{-3}	$(8.82\pm14.34)\times10^{-4}$	$(6.30\pm10.58)\times10^{-3}$	$(1.51\pm1.51)\times10^{-3}$
BiCoGAN	1.07×10^{-2}	$(2.77\pm0.41)\times10^{-3}$	$(7.28\pm1.11)\times10^{-3}$	$(2.32\pm0.31)\times10^{-2}$
Tsirtsis	2.05×10^{-4}	$(7.83\pm0.15)\times10^{-5}$	$(2.29\pm0.12)\times10^{-4}$	$(3.01\pm0.17)\times10^{-4}$

Reacher. Table S5 shows that CACTI has the lowest all-query, high-support, and low-direct-support errors, while Gaussian is only marginally lower in the medium region. CACTI is stable across seeds, whereas Conditional Flow shows substantial variation.

Table S5: Reacher normalized CF-MSE across direct-support regions. Uncertainties are sample standard deviations over model seeds $0, 1, 2$; lower is better.

Method	All	High	Medium	Low
CACTI	$\mathbf{4.07\times10^{-5}}$	$\mathbf{(1.281 \pm 0.004)\times10^{-5}}$	$(3.280 \pm 0.010)\times10^{-5}$	$\mathbf{(7.920 \pm 0.026)\times10^{-5}}$
Base-D	1.14×10^{-4}	$(4.73 \pm 1.20)\times10^{-5}$	$(7.46 \pm 1.40)\times10^{-5}$	$(2.35 \pm 0.09)\times10^{-4}$
Gaussian	6.48×10^{-5}	$(1.918 \pm 0.004)\times10^{-5}$	$\mathbf{(3.254 \pm 0.014)\times10^{-5}}$	$(1.534 \pm 0.001)\times10^{-4}$
Base-M	1.20×10^{-4}	$(3.70 \pm 0.71)\times10^{-5}$	$(6.73 \pm 0.67)\times10^{-5}$	$(2.726 \pm 0.009)\times10^{-4}$
Conditional Flow	1.15×10^{-3}	$(9.28 \pm 8.51)\times10^{-5}$	$(1.66 \pm 2.70)\times10^{-3}$	$(1.52 \pm 1.16)\times10^{-3}$
BiCoGAN	9.04×10^{-2}	$(3.050 \pm 0.051)\times10^{-2}$	$(6.990 \pm 0.101)\times10^{-2}$	$(1.777 \pm 0.027)\times10^{-1}$
Tsirtsis	1.86×10^{-3}	$(2.171 \pm 0.048)\times10^{-4}$	$(1.223 \pm 0.089)\times10^{-3}$	$(4.352 \pm 0.128)\times10^{-3}$

E.2 State-Block and Semantic Diagnostics

MountainCar11. Table S6 reports semantic and reconstruction diagnostics. Near-zero factual reconstruction for the invertible models verifies numerical inverse–forward consistency, but it does not guarantee correct nonfactual pairing. CACTI also has the smallest action-effect discrepancy and no observed reward or goal flip in this evaluation.

Table S6: MountainCar11 semantic and reconstruction diagnostics under the factual event representation. Lower is better.

Method	Reward err.	Goal flip	Action-effect err.	Factual MSE
CACTI	**0**	**0**	$\mathbf{3.90\times10^{-9}}$	7.46×10^{-17}
CACTI-S	0	0	3.76×10^{-8}	1.07×10^{-16}
Gaussian	5.33×10^{-3}	5.33×10^{-5}	1.42×10^{-7}	1.96×10^{-15}
Conditional Flow	0	0	6.61×10^{-7}	9.50×10^{-15}
CACTI-M	0	0	1.24×10^{-8}	8.76×10^{-4}
BiCoGAN	3.95×10^{-1}	3.95×10^{-3}	7.75×10^{-4}	2.17×10^{-5}
Tsirtsis	6.97×10^{-2}	6.97×10^{-4}	1.43×10^{-6}	1.97×10^{-15}

PointMaze. Table S7 shows that velocity is the main PointMaze error source. The no-mask CACTI model is slightly more accurate than CACTI-S on the dynamic state and preserves the desired goal. Post-hoc goal repair removes BiCoGAN's PointMaze goal-coordinate violation but leaves its non-goal error at 3.99×10^{-2}.

Table S7: PointMaze state-block and semantic errors. Uncertainties, where reported, are sample standard deviations over model seeds $0, 1, 2$; lower is better.

Method	Position	Velocity	Goal	Reward err.	Success flip
CACTI-S	1.99×10^{-7}	5.63×10^{-5}	0	6.93×10^{-5}	9.60×10^{-5}
CACTI	$\mathbf{1.72\times10^{-8}}$	$\mathbf{5.13\times10^{-5}}$	**0**	$\mathbf{1.83\times10^{-5}}$	$\mathbf{3.26\times10^{-5}}$
CACTI-M	2.55×10^{-5}	1.88×10^{-3}	1.52×10^{-24}	9.19×10^{-4}	1.42×10^{-3}
Gaussian	1.03×10^{-6}	3.63×10^{-3}	0	1.43×10^{-4}	2.44×10^{-4}
Conditional Flow	1.96×10^{-4}	3.03×10^{-2}	2.76×10^{-20}	1.00×10^{-4}	1.45×10^{-4}
BiCoGAN	1.24×10^{-2}	1.21×10^{-2}	6.49×10^{-3}	2.98×10^{-2}	2.62×10^{-2}

Reacher. Joint velocity dominates the state-block error in Table S8, and CACTI is lowest in every reported block.

Table S8: Reacher state-block errors. Uncertainties are sample standard deviations over model seeds $0, 1, 2$; lower is better.

Method	Cos/Sin [0:4)	Target [4:6)	Joint velocity [6:8)	Fingertip–target [8:10)
CACTI	$\mathbf{(1.277 \pm 0.020) \times 10^{-7}}$	$\mathbf{0}$	$\mathbf{(1.261 \pm 0.004) \times 10^{-3}}$	$\mathbf{(9.874 \pm 0.244) \times 10^{-9}}$
Base-D	$(2.141 \pm 0.179) \times 10^{-6}$	$(1.79 \pm 1.78) \times 10^{-25}$	$(3.403 \pm 0.360) \times 10^{-3}$	$(1.104 \pm 0.090) \times 10^{-7}$
Gaussian	$(6.984 \pm 0.080) \times 10^{-7}$	$(3.37 \pm 2.73) \times 10^{-25}$	$(1.968 \pm 0.002) \times 10^{-3}$	$(4.069 \pm 0.095) \times 10^{-8}$
Base-M	$(3.123 \pm 0.412) \times 10^{-6}$	$(1.16 \pm 1.87) \times 10^{-23}$	$(3.475 \pm 0.144) \times 10^{-3}$	$(1.580 \pm 0.278) \times 10^{-7}$
Conditional Flow	$(7.000 \pm 5.865) \times 10^{-5}$	$(7.59 \pm 13.15) \times 10^{-23}$	$(5.391 \pm 1.837) \times 10^{-3}$	$(4.957 \pm 6.375) \times 10^{-5}$
BiCoGAN	$(2.261 \pm 1.020) \times 10^{-3}$	$(6.492 \pm 3.233) \times 10^{-5}$	(2.656 ± 0.042)	$(9.050 \pm 0.874) \times 10^{-5}$
Tsirtsis	$(1.522 \pm 0.051) \times 10^{-4}$	$(6.04 \pm 6.49) \times 10^{-25}$	$(4.213 \pm 0.127) \times 10^{-2}$	$(1.031 \pm 0.002) \times 10^{-5}$

E.3 Event-Matching Perturbation Diagnostics

Factual U uses the event representation inferred from the matching factual transition. Shuffled U swaps representations across factual events. Fixed U reuses one latent value across queries; $g_\psi(s, u)$ remains state-dependent. These are evaluation modes of a trained model, with the factual state, query action, and shared response held fixed. The shuffled and fixed modes perturb factual-event matching, as formalized by Eq. (S32) in Appendix B.5.

Figure 3 reports the CACTI perturbation diagnostic. Table S9 reports the detailed Tsirtsis results: shuffled and fixed event representations both increase its mean counterfactual error. The columns E_{fact}, E_{shuf}, and E_{fixed} report factual, shuffled, and fixed-mode errors; R_{shuf} and R_{fixed} report their multipliers relative to factual-mode error.

Table S9: Tsirtsis event-matching perturbation diagnostics. Low denotes the low-direct-support subset. Uncertainties are sample standard deviations over model seeds $0, 1, 2$.

Env.	Subset	E_{fact}	E_{shuf}	E_{fixed}	R_{shuf}	R_{fixed}
MC11	All	$(1.20 \pm 0.17) \times 10^{-4}$	$(5.717 \pm 0.013) \times 10^{-3}$	$(4.068 \pm 0.008) \times 10^{-3}$	48.24 ± 6.34	34.33 ± 4.57
MC11	Low	$(2.45 \pm 0.28) \times 10^{-4}$	$(5.847 \pm 0.007) \times 10^{-3}$	$(4.176 \pm 0.010) \times 10^{-3}$	24.03 ± 2.62	17.16 ± 1.90
PointMaze	All	$(2.053 \pm 0.095) \times 10^{-4}$	$(9.427 \pm 0.049) \times 10^{-4}$	$(5.630 \pm 0.052) \times 10^{-4}$	4.60 ± 0.19	2.75 ± 0.11
PointMaze	Low	$(3.013 \pm 0.171) \times 10^{-4}$	$(1.142 \pm 0.012) \times 10^{-3}$	$(6.650 \pm 0.142) \times 10^{-4}$	3.80 ± 0.17	2.21 ± 0.08

E.4 Exploratory Variants

CACTI-S uses the structural-mask implementation described in Appendix D.1. Its exploratory comparisons appear in Tables S2, S4, S6, and S7, alongside the primary no-mask CACTI model. The low-direct-support entries in Tables S2 and S4 contain the values underlying Figure 2.

F Downstream Policy Learning and Scope

F.1 Full Downstream Protocol

Figure 4 of the main paper reports the final outcomes. MountainCar11 and PointMaze use $N_{\text{real}} \in \{500, 600, 750, 1000, 1200\}$, a synthetic-to-real ratio of 0.5, and policy seeds $0, 1, 2$. All methods share the same factual subset, and augmented methods share the same alternative-action queries. MountainCar11 uses return; PointMaze uses success rate and selects checkpoints with a goal-reaching validation criterion.

Related offline-RL work includes the D4RL benchmark suite (Fu et al., 2021), model-free policy learning with behavior-cloning regularization (Fujimoto & Gu, 2021), and conservative model-based policy optimization (Yu et al., 2021). Data augmentation has also been studied with causal models

(Zhang et al., 2026), causal normalizing flows (Cho & Sun, 2025), and guided generative trajectory models (Lee et al., 2024).

F.2 Scope and Limitations

Sequential exogeneity can fail when the behavior action anticipates an unrecorded next-step disturbance. The action-independent exogenous mechanism also restricts strong action–disturbance interactions, and the shared response is only a local approximation. Directions outside $\mathrm{Range}(\Sigma_A(s))$ are not identified without additional restrictions, while large interventions may incur nonlinear approximation error. The learned event coordinates need not equal the physical disturbance coordinates; the simulator oracle is used only for evaluation. This work studies one-step counterfactuals, and RotateZ shows that high-dimensional same-event inference remains challenging.

Related counterfactual applications include explanations for sequential decisions (Tsirtsis et al., 2021), temporal point processes (Noorbakhsh & Rodriguez, 2022), and harm assessment (Richens et al., 2022).