CACTI: Control-Structured Causal Abduction for Counterfactual Transition Inference in Offline Reinforcement Learning

Jingyang He, Shuai Wang, Zhen Lei, Ang Li*

Department of Computer Science, Florida State University
jh24o@fsu.edu, sw23v@fsu.edu, zl24a@fsu.edu, angli@cs.fsu.edu

Abstract

Counterfactual transition inference asks what next state would have occurred under a different action while preserving the exogenous realization U of the observed event. Predictive fit alone does not ensure this same-event correspondence, because a learned model may explain the same factual transition by trading off action effects and latent exogenous variation. We introduce CACTI (Causal Abduction for Counterfactual Transition Inference), a control-structured transition SCM that separates action-independent evolution, a state-dependent response shared across actions, and an invertible exogenous mechanism that does not directly receive the action. CACTI abducts $\widehat{U}$ from the factual transition, replaces the action, and predicts with the same realization. This structure reduces action–U ambiguity and allows related action variation to constrain weakly supported queries. Our analysis characterizes the local validity and action-direction coverage of the shared response. Using a same-U simulator oracle across four control tasks, CACTI improves counterfactual fidelity over predictive and explicit-abduction baselines, including BiCoGAN and the Tsirtsis location–scale SCM. Ablations, U-perturbation diagnostics, and low-data policy experiments further verify the proposed mechanisms and downstream utility.

Introduction

Counterfactual transition inference asks what would have happened to an observed transition under a different action while preserving the same exogenous realization U. This is different from ordinary dynamics prediction. Given (s_t, a_t, s_{t+1}), a predictive model only needs to describe plausible outcomes under a'_t, whereas a same-event counterfactual must preserve the correspondence between the factual and alternative-action worlds (Peters, Janzing, and Schölkopf 2017; Oberst and Sontag 2019).

Consider a vehicle that turns right under a particular crosswind and then drifts off the road. The observed transition records only the combined effect of steering and wind; it does not reveal how much of the displacement came from each. Two models may reconstruct the factual next state equally well while assigning different effects to the action and the crosswind. After replacing the right turn with a left turn, these models will reuse different interpretations of the same event and can therefore produce different counterfactual outcomes.

A transition SCM represents unobserved influences through

$$S_{t+1} = f^\star(S_t, A_t, U_{t+1}), \qquad S_{t+1}(a'_t) = f^\star(s_t, a'_t, u_{t+1}).$$

The realization u_{t+1}, which captures event-specific influences such as the particular gust of crosswind, is inferred from the factual transition and kept fixed when the action is replaced. For a flexible learned SCM $f_\theta(S, A, U)$, however, factual data may not uniquely determine how the observed state change should be divided between the action response and U. Accurate factual reconstruction therefore does not guarantee that the inferred U retains a stable meaning across actions (Nasr-Esfahany and Kiciman 2023).

Existing causally motivated methods fall into two broad groups. Some use causal structure, modularity, or counterfactual constraints to guide augmentation or filter model rollouts without explicitly preserving an abducted factual realization (Pitis, Creager, and Garg 2020; Zhang et al. 2026; Cho and Sun 2025). Explicit-U transition SCMs instead recover the factual realization and reuse it after intervention (Lu et al. 2020; Tsirtsis and Rodriguez 2023). CACTI retains this same-event semantics while additionally separating a response shared across actions from event-specific exogenous variation, reducing the action–U ambiguity of flexible action-conditioned mechanisms.

Both groups are often evaluated through task-level gains, but such gains do not establish that an individual transition is the correct counterfactual; they may instead reflect changes in coverage, regularization, or replay composition. We therefore evaluate same-U counterfactual fidelity directly and use downstream policy learning as a separate test of augmentation utility.

The distinction between counterfactual correctness and downstream utility is especially relevant for the actions that motivate augmentation: those for which direct experience is scarce. We call this regime *weak support*: the queried action itself is sparsely observed, but nearby or directionally related action variations remain in the data. Such a query is neither well covered nor unconstrained; it has little direct evidence, while related actions may still reveal the response

*Corresponding author.

direction needed by the query. This setting tests whether a model can use cross-action evidence without changing the event-specific meaning of U. If the relevant action direction is absent altogether, the query cannot be recovered from observational data alone.

We assume sequential exogeneity, $U_{t+1} \perp A_t \mid S_t$, meaning that, given the recorded state, the behavior action is not selected in anticipation of the disturbance realized during the next transition. This is natural when S_t contains the information used by the behavior policy and U_{t+1} represents subsequent process noise, as in standard counterfactual-RL transition SCMs (Lu et al. 2020; Tsirtsis and Rodriguez 2023). Within this setting, we introduce CACTI (*Causal Abduction for Counterfactual Transition Inference*), which uses the control-inspired decomposition

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

The three terms distinguish what follows predictably from the current state, what changes systematically with the action, and what is specific to the realized event. In the vehicle example, $m_\theta(S_t)$ captures motion implied by the current velocity and heading, $B_\theta(S_t)A_t$ captures the steering response, and $g_\psi(S_t, U_{t+1})$ captures the particular unrecorded gust of crosswind. Here, U indexes influences not determined by the modeled state and action.

By separating event-specific exogenous variation from a response shared across actions, CACTI preserves same-event alignment while using observed action directions to constrain weak-support queries.

Our analysis characterizes the shared response as a local approximation and identifies the action directions supported by the data. Empirically, a same-U oracle evaluates fidelity, ablations and U-perturbations isolate the mechanism, and low-data policy experiments test downstream utility. Across four environments, CACTI outperforms predictive and explicit-abduction baselines, including BiCoGAN and the Tsirtsis location–scale SCM, with consistent gains under weak support.

Contributions. **(1)** We identify the gap between factual dynamics fit and same-event cross-action correspondence under uneven action coverage. **(2)** We introduce CACTI, a control-structured SCM that separates action-independent evolution, a shared action response, and event-specific exogenous variation. **(3)** We develop a same-U simulator-oracle protocol and evaluate fidelity, weak-support robustness, the role of inferred U, and downstream utility across diverse control tasks.

Preliminaries

We consider a factual one-step transition

$$e_t = (s_t, a_t, s_{t+1})$$

from an offline reinforcement-learning dataset. Its transition mechanism is represented by a structural causal model (Peters, Janzing, and Schölkopf 2017):

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

Here, U_{t+1} is exogenous: its causes are not modeled within the structural equations, and it represents influences on S_{t+1} not explicitly included in S_t and A_t, such as the particular gust of crosswind in the introduction.

Assuming that $f^\star_{s,a}(u) := f^\star(s, a, u)$ is bijective in u (Nasr-Esfahany, Alizadeh, and Shah 2023), Pearl-style counterfactual reasoning proceeds in three steps:

1. **Abduction:** recover the factual exogenous realization as $u_{t+1} = (f^\star_{s_t,a_t})^{-1}(s_{t+1})$.
2. **Intervention:** replace the factual action by $A_t \leftarrow a'_t$, while keeping s_t and u_{t+1} fixed.
3. **Prediction:** evaluate the intervened mechanism to obtain

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

Keeping the same u_{t+1} couples the factual and counterfactual outcomes at the level of the observed event (Pawlowski, Coelho de Castro, and Glocker 2020). Equation (3) is the counterfactual transition studied throughout this paper.

CACTI: Separating Action Effects from Factual Disturbances

Why Predictive Fit Is Not Enough

An expressive transition model can accurately fit the outcome distribution under every action. A same-event counterfactual requires more: it must determine which outcomes under different actions correspond to the same exogenous realization.

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.

Proposition 1 *Identical action-conditional outcome distributions do not, in general, determine the same-event counterfactual.*

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. Thus, accurately fitting a transition or residual distribution does not by itself give a latent variable a consistent exogenous meaning across actions (Nasr-Esfahany and Kiciman 2023; Chen and Du 2025). CACTI addresses this ambiguity by placing action-dependent variation in a control-response mechanism shared across actions, while using an exogenous mechanism that does not directly depend on the action.

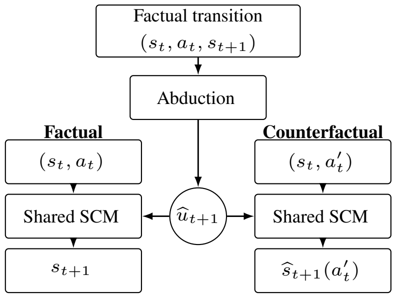

Figure 1: CACTI training pipeline. The shared response and inferred U jointly reconstruct the factual transition; dashed arrows denote training constraints.

A Control Structure for Counterfactuals

The crosswind example motivates the two key structural choices in Eq. (1). The exogenous mechanism conditions on S_t because the same disturbance can have different effects at different speeds, headings, or system configurations. It deliberately excludes A_t, so that replacing the action does not also change the interpretation of the abducted U; otherwise, an action-conditioned decoder could reassign part of the action effect to the exogenous representation. Systematic action dependence is instead isolated in the shared response $B_\theta(S_t)A_t$. This is a structural restriction rather than a globally linear dynamics assumption: m_θ, B_θ, and g_ψ may all be nonlinear, while the action enters only through the state-dependent response $B_\theta(S_t)A_t$.

For brevity, define

$$\mu_\theta(s, a) := m_\theta(s) + B_\theta(s)a.$$

For a fixed state s, we write $g_{\psi,s}(u) := g_\psi(s, u)$ and assume that this map is invertible (Pawlowski, Coelho de Castro, and Glocker 2020; Nasr-Esfahany, Alizadeh, and Shah 2023). Given a factual transition (s_t, a_t, s_{t+1}) and a counterfactual action a'_t, CACTI follows three steps.

Abduction. Infer the exogenous realization associated with the factual transition:

$$\widehat{u}_{t+1} = g^{-1}_{\psi,s_t}\left(s_{t+1} - s_t - \mu_\theta(s_t, a_t)\right). \tag{4}$$

The quantity inside the inverse is the realized effect of the exogenous variable on the state change, rather than U_{t+1} itself.

Intervention. Replace the factual action by $A_t \leftarrow a'_t$, while keeping s_t and $\widehat{u}_{t+1}$ fixed.

Prediction. Evaluate the transition under the counterfactual action:

$$\widehat{s}_{t+1}(a'_t) = s_t + \mu_\theta(s_t, a'_t) + g_\psi(s_t, \widehat{u}_{t+1}) \tag{5}$$

The factual next state is used only to abduct the event-specific realization $\widehat{u}_{t+1}$, not as a baseline for the counterfactual outcome. CACTI reevaluates the learned structural equation at a'_t while keeping $(s_t, \widehat{u}_{t+1})$ fixed, so the factual transition determines the shared exogenous realization and the query determines the new control response.

Because $g_\psi(S_t, U_{t+1})$ does not receive the action while $B_\theta(S_t)$ is shared across actions, CACTI excludes the action-dependent relabeling in Section and allows other actions to constrain a less-observed query. The next subsection characterizes the local validity of this response and the action-direction coverage required by the query.

Learning CACTI from Offline Transitions

Let

$$\mathcal{D} = \{(s_t, a_t, s_{t+1})\}_{t=1}^N$$

denote the offline transition dataset. It contains neither the realized exogenous variable u_{t+1} nor counterfactual outcomes under alternative actions. Consequently, abduction, intervention, and prediction are not separately supervised during training. CACTI instead jointly estimates m_θ, B_θ, and the invertible exogenous mechanism g_ψ using an objective constructed entirely from factual transitions. Once fitted, invertibility makes the event-specific realization recoverable from an observed transition and reusable when evaluating an alternative action.

Learning the factual transition SCM. For each factual transition (s_t, a_t, s_{t+1}), we work in state-increment space and define

$$\Delta s_{t+1} := s_{t+1} - s_t, \qquad \mu_\theta(s_t, a_t) := m_\theta(s_t) + B_\theta(s_t)a_t.$$

Here, μ_θ is shorthand for the systematic state- and action-dependent component of Eq. (1). This notation separates the part predicted from (s_t, a_t) from the remaining event-specific contribution modeled by the invertible exogenous mechanism.

The event-specific exogenous effect is represented by a state-conditioned invertible mechanism $g_\psi(s_t, u_{t+1})$. The action enters only through $\mu_\theta(s_t, a_t)$ and is not provided to the exogenous mechanism. The resulting transition SCM is

$$\Delta S_{t+1} = \mu_\theta(S_t, A_t) + g_\psi(S_t, U_{t+1}). \tag{6}$$

Figure 2 summarizes the corresponding learning pipeline.

Equation (6) induces a conditional density over the observed state increment,

$$p_{\theta,\psi}\left(\Delta S_{t+1} \mid S_t, A_t\right).$$

We train this conditional density by change of variables, as in normalizing flows (Dinh, Sohl-Dickstein, and Bengio 2017; Khemakhem et al. 2021), using the factual negative log-likelihood:

$$\mathcal{L}_{\text{fact}} = -\mathbb{E}_{(s_t,a_t,s_{t+1})\sim\mathcal{D}}\left[\log p_{\theta,\psi}\left(\Delta s_{t+1} \mid s_t, a_t\right)\right]. \tag{7}$$

This objective does not require labels for the true u_{t+1}. The invertible mechanism learns an exogenous representation for the event-specific variation observed in the data. Its reference density only fixes the numerical coordinates of the learned exogenous variable; it is not an assumption that the physical disturbance follows that distribution. The exact likelihood calculation is provided in the appendix.

The factual likelihood alone requires the shared action mechanism and the exogenous mechanism to jointly explain

Algorithm 1: CACTI Training and Same-Event Inference

Require: Offline transitions $\mathcal{D}$, iterations K
$\mu_\theta(s,a) \leftarrow m_\theta(s) + B_\theta(s)\phi(a)$
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: (s_t, a_t, s_{t+1}) and a'_t
$z^\theta_{t+1} \leftarrow (s_{t+1} - s_t) - \mu_\theta(s_t, a_t)$
$\widehat{u}_{t+1} \leftarrow g^{-1}_{\psi, s_t}(z^\theta_{t+1})$ ▷ Abduction
Keep $(s_t, \widehat{u}_{t+1})$ fixed and set $A_t \leftarrow a'_t$
$\widehat{s}_{t+1}(a'_t) \leftarrow s_t + \mu_\theta(s_t, a'_t) + g_\psi(s_t, \widehat{u}_{t+1})$
return $\widehat{s}_{t+1}(a'_t)$

the observed transition. To provide a direct learning signal for the shared response, we additionally use

$$\mathcal{L}_{\text{resp}} = \mathbb{E}_{\mathcal{D}}\left[\|\Delta s_{t+1} - m_\theta(s_t) - B_\theta(s_t)a_t\|_2^2\right]. \quad (8)$$

This auxiliary objective encourages systematic action-dependent variation to be represented by m_θ and B_θ. It does not assume deterministic dynamics: event-specific variation that is not explained by the shared mechanism remains modeled by g_ψ.

We also regularize the magnitude of the shared response:

$$\mathcal{R}_B = \mathbb{E}_{\mathcal{D}}\left[\|B_\theta(s_t)\|_F^2\right]. \quad (9)$$

This term discourages unrealistically large action sensitivity when the offline data are finite or unevenly distributed across actions.

The complete learning objective is

$$\mathcal{L}_{\text{CACTI}} = \mathcal{L}_{\text{fact}} + \lambda_{\text{resp}}\mathcal{L}_{\text{resp}} + \lambda_B \mathcal{R}_B. \quad (10)$$

Because the factual transition density depends on both the shared mechanism and Because the factual likelihood couples the shared and exogenous mechanisms, optimizing Eq. (10) jointly constrains m_θ, B_θ, and g_ψ. We first stabilize m_θ and B_θ with the response and regularization terms, then optimize all components jointly; normalization, architectures, and environment-specific schedules are detailed in the appendix.

Counterfactual generation. After training, CACTI applies the learned SCM to a factual transition. For compactness, write $g_{\psi,s}(u) := g_\psi(s,u)$. Using $\Delta s_{t+1} = s_{t+1} - s_t$, CACTI first infers the event-specific exogenous realization and then evaluates the same realization under the counterfactual action:

$$\begin{aligned} \widehat{u}_{t+1} &= g^{-1}_{\psi,s_t}\left(\Delta s_{t+1} - \mu_\theta(s_t, a_t)\right), \\ \widehat{s}_{t+1}(a'_t) &= s_t + \mu_\theta(s_t, a'_t) + g_\psi(s_t, \widehat{u}_{t+1}). \end{aligned} \quad (11)$$

The first line uses the factual transition to identify the exogenous realization associated with that event. The second replaces a_t with the prescribed action a'_t while keeping the same $\widehat{u}_{t+1}$. No new exogenous realization is sampled after the action is changed.

The next section analyzes when the shared response is locally appropriate and when the offline data contain sufficient action information for the query.

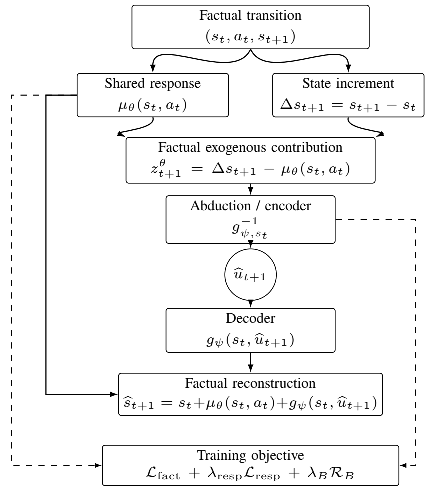

Figure 2: CACTI training pipeline. The shared response models systematic action effects, while the event-specific contribution is encoded as $\widehat{u}_{t+1}$ and decoded for factual reconstruction. Dashed arrows denote training constraints.

Structural Analysis under Weak Support

Relative to a generic learned transition SCM, CACTI imposes two cross-action restrictions: a shared exogenous mechanism $g_\psi(S_t, U_{t+1})$, which preserves the meaning of an abducted realization after intervention (Oberst and Sontag 2019; Tsirtsis, De, and Rodriguez 2021; Noorbakhsh and Rodriguez 2022), and a shared response $B_\theta(S_t)$, which lets better-supported actions inform weakly observed queries. We analyze the resulting alignment, the response's local approximation error, and the action-direction coverage required by a query; proofs are in the appendix.

Cross-Action Alignment and Local Validity

Cross-action alignment. Proposition 1 locates the ambiguity in an action-specific change of latent coordinates. A generic generator $G(s,a,u)$ can use one relabeling of u under action a and another under a', while preserving both action-conditional transition distributions. Such relabelings change which exogenous realization in the counterfactual world is treated as corresponding to the factual event (Nasr-Esfahany and Kiciman 2023; Chen and Du 2025).

CACTI rules out this specific degree of freedom by using a single invertible exogenous map $g_{\psi,s} : u \mapsto g_\psi(s,u)$ for all actions, building on the tractable inversion available in bijective and flow-based SCMs (Pawlowski, Coelho de Castro, and Glocker 2020; Khemakhem et al. 2021; Nasr-Esfahany, Alizadeh, and Shah 2023). An intervention changes the systematic response $\mu_\theta(s,a)$, but not the latent coordinate system used by the exogenous mechanism. An action-independent reparameterization of U may still remain, but because it is shared across action worlds, it preserves their event-level correspondence (Chen and Du 2025). Thus, cross-action alignment refers to preservation of the same-event pairing, not numerical recovery of the physical disturbance.

Explicit factual-event correspondence. For a learned CACTI model, define

$$\widehat{F}_{\theta,\psi}(s, a; u) := s + \mu_\theta(s, a) + g_\psi(s, u). \tag{12}$$

Given a factual transition $e_t = (s_t, a_t, s_{t+1})$, CACTI abducts the exogenous realization through the invertible transition mechanism (Pawlowski, Coelho de Castro, and Glocker 2020; De Sousa Ribeiro et al. 2023; Tsirtsis and Rodriguez 2023):

$$\widehat{u}_{t+1} = g^{-1}_{\psi,s_t}\left(s_{t+1} - s_t - \mu_\theta(s_t, a_t)\right). \tag{13}$$

For any candidate exogenous realization $\widetilde{u}$, define its *factual-consistency discrepancy*

$$c^{\mathrm{F}}_t(\widetilde{u}) := \widehat{F}_{\theta,\psi}(s_t, a_t; \widetilde{u}) - s_{t+1}. \tag{14}$$

This quantity does not define U as a prediction residual. Rather, it measures whether a proposed exogenous realization is consistent with the observed factual event under the learned SCM.

By exact inversion,

$$c^{\mathrm{F}}_t(\widetilde{u}) = g_\psi(s_t, \widetilde{u}) - g_\psi(s_t, \widehat{u}_{t+1}). \tag{15}$$

Because g_{ψ,s_t} is injective,

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

Thus, $\widehat{u}_{t+1}$ is the unique realization, in the learned exogenous coordinate system, that matches the factual transition, as in bijective-SCM abduction (Nasr-Esfahany, Alizadeh, and Shah 2023; Tsirtsis and Rodriguez 2023). This does not require its coordinates to numerically coincide with those of the physical disturbance (Chen and Du 2025).

Now consider a query action a'_t, let $d_t := a'_t - a_t$, and define the true same-event action contrast

$$\tau^\star_t(a'_t) := F^\star(s_t, a'_t, u_{t+1}) - F^\star(s_t, a_t, u_{t+1}). \tag{16}$$

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

$$\begin{aligned} &\widehat{F}_{\theta,\psi}(s_t, a'_t; \widetilde{u}) - F^\star(s_t, a'_t, u_{t+1}) \\ &\quad = c^{\mathrm{F}}_t(\widetilde{u}) + B_\theta(s_t) d_t - \tau^\star_t(a'_t). \end{aligned} \tag{17}$$

The decomposition separates factual-event correspondence from action-response estimation. The realization $\widetilde{u}$ determines whether the prediction remains matched to the factual event, whereas $B_\theta(s_t)d_t$ determines how that event responds to the intervention. For $\widetilde{u} = \widehat{u}_{t+1}$, the factual-consistency term vanishes, leaving only the shared-response error. Exact inversion therefore yields an algebraically reduced matched prediction, but this is only the matched-realization case of the explicit SCM. Keeping U explicit exposes factual-event correspondence as a model variable that can be perturbed while the state, query, and shared response remain fixed (Kazemi et al. 2025; Lally, Kazemi, and Paoletti 2026).

For any a'_t,

$$\widehat{F}_{\theta,\psi}(s_t, a'_t; \widetilde{u}) - \widehat{F}_{\theta,\psi}(s_t, a'_t; \widehat{u}_{t+1}) = c^{\mathrm{F}}_t(\widetilde{u}). \tag{18}$$

Thus, shuffled or fixed realizations isolate changes in factual-event correspondence and motivate our U-perturbation diagnostics. A perturbation may occasionally offset another model error, so degradation is not guaranteed pointwise; systematic degradation across queries indicates that accurate prediction depends on preserving the matched factual realization.

Why share the response? The event-specific action Jacobian

$$J_A F^\star(s_t, a_t, u_{t+1})$$

depends on an unobserved factual realization, while each factual transition reveals only one action–outcome pair for that event. Without additional cross-action identifying restrictions, factual fit alone does not determine a response that is allowed to vary freely with (S, A, U) at an alternative action (Nasr-Esfahany and Kiciman 2023; Nasr-Esfahany, Alizadeh, and Shah 2023; Immer et al. 2023; Tsirtsis and Rodriguez 2023). CACTI instead pools action variation through a common $B_\theta(S_t)$, allowing better-supported actions to inform the response required by a weakly supported query. The approximation cost appears as $B_\theta(s_t)d_t - \tau^\star_t(a'_t)$ in Proposition 2.

Action Coverage under Weak Support

To isolate action coverage from response-approximation error, suppose that the shared-response model is exact in the local region of interest:

$$\Delta S_{t+1} = m^\star(S_t) + B^\star(S_t)A_t + g^\star(S_t, U_{t+1}). \tag{19}$$

At a fixed state s, we treat $B^\star(s)$ nonparametrically, without additional smoothness or cross-state parameter sharing.

We assume sequential exogeneity,

$$U_{t+1} \perp A_t \mid S_t, \tag{20}$$

meaning that, given the recorded state, the behavior action does not anticipate the disturbance realized in the next transition. This separates observed action variation from disturbance selection; it does not require every factual event to have exactly the same local action response (Lu et al. 2020; Tsirtsis and Rodriguez 2023).

Let $A_t \in \mathbb{R}^{d_a}$ denote the numerical control coordinates used by the transition model. For a state s, define

$$\begin{aligned} \Sigma_A(s) &:= \mathrm{Cov}\left(A_t \mid S_t = s\right), \\ C_{\Delta A}(s) &:= \mathrm{Cov}\left(\Delta S_{t+1}, A_t \mid S_t = s\right). \end{aligned} \tag{21}$$

The range of $\Sigma_A(s)$ records the centered action directions that vary in the observational distribution around state s. For continuous states, these are regular conditional moments defined for P_{S_t}-almost every s; finite-sample estimation relies on local neighborhoods or function approximation.

Theorem 1 (Action-direction coverage) *Under Eqs. (19) and (20), assume the relevant conditional second moments exist and let $d_t = a'_t - a_t$. Within the state-local shared-response model, and without additional cross-state restrictions, the query effect*

$$B^\star(s_t)d_t$$

is uniquely determined by the observational distribution if and only if

$$d_t \in \mathrm{Range}\left(\Sigma_A(s_t)\right). \tag{22}$$

When this condition holds,

$$B^\star(s_t)d_t = C_{\Delta A}(s_t)\Sigma_A(s_t)^\dagger d_t, \tag{23}$$

where † denotes the Moore–Penrose pseudoinverse.

Sequential exogeneity implies

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

and hence

$$C_{\Delta A}(s) = B^\star(s)\Sigma_A(s). \tag{24}$$

If d_t lies in the range of $\Sigma_A(s_t)$, every response matrix consistent with Eq. (24) agrees on $B^\star(s_t)d_t$. If d_t has a component in the null space of $\Sigma_A(s_t)$, the response along that component can be changed without altering the observed conditional moments. The complete proof is provided in the appendix.

For example, if vehicle controls include steering and acceleration but only steering varies in the offline data, local steering effects are identifiable whereas acceleration effects are not. The exact steering value need not have appeared, provided its change lies in the observed span; this condition concerns direction rather than magnitude, so large changes may still incur nonlinear approximation error. More generally, $\Sigma_A(s) \succ 0$ identifies all local directions, while a singular covariance identifies only its range; weakly explored directions may remain identifiable but be unstable in finite samples.

Experiments

Experimental setup. Our evaluation follows a correctness–mechanism–utility progression: event-level same-U fidelity (Richens, Beard, and Thompson 2022); whether gains arise from the shared response and matched factual U, rather than capacity or latent abduction alone; and whether the resulting transitions improve fixed-dataset offline policy learning with limited data (Fu et al. 2021). We study MountainCar11, PointMaze, Reacher, and Shadow Hand RotateZ, spanning ordered discrete, goal-conditioned, nonlinear continuous, and high-dimensional control.

A same-U oracle restores each factual simulator state, reuses its realized disturbance, changes only the action, and executes one step. The oracle U is used only to construct targets and is never given to the learned models. We induce uneven coverage through nonuniform behavior policies; direct support is measured by empirical frequency in MountainCar11 and action density in the continuous environments, so weak-support bins retain related action variation despite sparse direct evidence.

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-U counterfactual fidelity. Each cell reports normalized CF-MSE on all nonfactual queries (top) and weak-support queries (bottom); lower is better.

We compare CACTI with Base-D, Gaussian dynamics, Base-M (Bishop 1994), Conditional Flow (Dinh, Sohl-Dickstein, and Bengio 2017), BiCoGAN (Lu et al. 2020), and the Tsirtsis location–scale SCM (Tsirtsis and Rodriguez 2023). Our primary metric is normalized CF-MSE between the predicted and oracle counterfactual next states, using the evaluator's environment-specific normalization. We report results over all nonfactual queries and the weak-support subset, using identical factual transitions and counterfactual actions across methods and averaging over three model seeds.

For downstream evaluation, MountainCar11 and PointMaze use $N_{\mathrm{real}} \in \{500, 600, 750, 1000, 1200\}$, a synthetic-to-real ratio of 0.5, and policy seeds `0,1,2`. All methods use the same factual subset, and augmented variants use the same alternative-action queries. Implementation details and additional diagnostics are provided in the appendix.

Counterfactual Fidelity across Environments

Table 1 compares CACTI with deterministic, probabilistic, flow-based, and explicit-abduction SCM baselines. CACTI achieves the lowest error in settings, improving over the strongest baseline by one to two orders of magnitude on MountainCar11 and PointMaze, with smaller but consistent gains on Reacher and RotateZ. The strongest non-CACTI method varies Gaussian on MountainCar11 and Reacher, Tsirtsis on PointMaze, and Gaussian and Base-M on the all-query and weak-support RotateZ columns, respectively, so the gains cannot be attributed to one distributional assumption, model capacity, or explicit U-abduction alone. CACTI also remains best under weak support, consistent with leveraging related action variation; this does not imply recovery along unobserved directions, and RotateZ shows that high-dimensional same-event inference remains challenging. Full support-stratified and state-block results are in the appendix.

Structural and Exogenous-Mechanism Analysis

Structural ablation. Figure 3 reports weak-support CF-MSE, where lower is better. The four controls isolate complementary explanations for CACTI's performance. CACTI-M retains the shared action response but removes the event-specific exogenous mechanism, testing whether an average

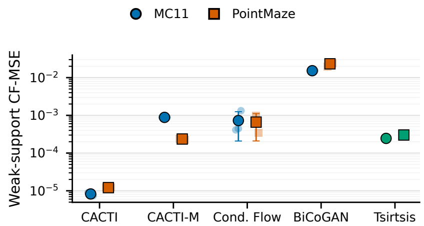

Figure 3: Structural ablation on MountainCar11 and PointMaze. We report weak-support CF-MSE; lower is better.

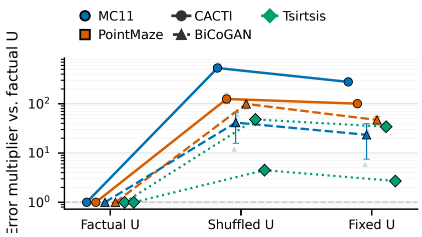

Figure 4: Inferred-U diagnostic. Larger multipliers under shuffled or fixed U indicate stronger degradation from disrupting event-specific correspondence.

response model is sufficient. Conditional Flow provides expressive conditional generation without CACTI's separation between action response and exogenous variation, controlling for model capacity. BiCoGAN and Tsirtsis both abduct and reuse U, but use adversarial and bijective location–scale mechanisms, respectively, without CACTI's shared action-response decomposition. CACTI substantially outperforms all four controls in both environments. The gain therefore cannot be explained solely by a mean response, a flexible generator, or the presence of an inferred exogenous variable.

Event-specific U diagnostic. Figure 4 tests the functional contribution of inferred $\widehat{U}$ to counterfactual accuracy. *Factual* U uses the realization inferred from its factual transition, not the simulator's true U; *Shuffled* U swaps realizations across transitions while preserving their empirical distribution but breaking the event–$\widehat{U}$ pairing; and *Fixed* U uses one realization for all transitions, removing event-to-event exogenous variation. The trained models, factual transitions, and queried actions remain unchanged.

Let E_m denote the counterfactual error under mode m. We report $R_m = E_m/E_{\text{fact}}$, so Factual U is normalized to one. Larger values under Shuffled or Fixed U indicate stronger degradation after the event–$\widehat{U}$ correspondence is broken, rather than worse fidelity under normal inference. Together with its lower absolute error, CACTI's pronounced degradation under both perturbations shows that its accuracy depends on preserving the exogenous realization inferred from the corresponding factual event. This provides a functional validation of the learned U.

Downstream Policy Utility under Limited Data

Figure 5 tests whether CACTI's higher-fidelity counterfactual transitions translate into useful offline-RL augmentation, as in model-based and generative augmentation work (Pitis et al. 2022; Lee et al. 2024). This question is separate from counterfactual correctness: downstream gains also depend on model error, replay composition, and offline optimization (Yu et al. 2020, 2021). We therefore compare real-only training with CACTI, Tsirtsis and BiCoGAN augmentation under limited real-data budgets.

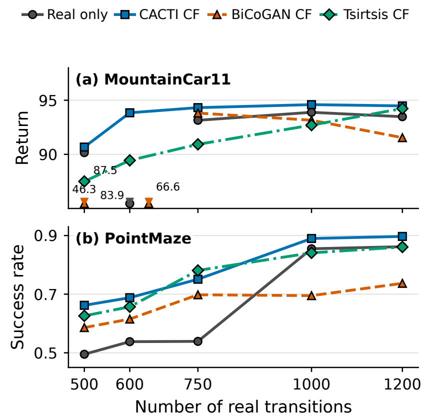

Figure 5: Downstream utility across five real-data budgets. Mean $\pm$ standard deviation over policy seeds `0,1,2`; higher is better.

On MountainCar11, CACTI achieves the highest mean return at every budget. At $N_{\text{real}} = 500$, it reaches 90.67, compared with 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. PointMaze shows a similar trend, with CACTI attaining the highest mean success rate at 500, 600, 1000, 1200. At $N_{\text{real}} = 1200$, CACTI achieves 0.897, compared with 0.862 for real-only training, 0.861 for Tsirtsis and 0.737 for BiCoGAN.

Across all ten settings, CACTI attains the highest mean primary metric in nine, supporting its utility for low-data augmentation. Together with the same-U fidelity and U-perturbation diagnostics, this supports that accurate counterfactual augmentation improves downstream policy learning (Kumar et al. 2020; Fujimoto and Gu 2021).

Conclusion

We studied same-event counterfactual transition inference in offline reinforcement learning, where alternative actions are evaluated while preserving the factual exogenous realization U. Flexible dynamics models can fit factual transitions yet entangle action effects with exogenous variation, especially under weak action support.

We introduced CACTI, a control-structured transition SCM that separates a shared state-dependent action response from event-specific exogenous variation and reuses the abducted U after intervention. Across four environments, CACTI improves same-U fidelity over predictive and explicit-abduction baselines, including BiCoGAN and the Tsirtsis location–scale SCM. Ablations, U-perturbations, and low-data policy experiments validate the decomposition and its downstream utility, supporting structured same-U augmentation under weak but informative action support.

References

Bishop, C. 1994. Mixture Density Networks. Technical Report NCRG/94/004.

Chen, Y.; and Du, D. 2025. Exogenous Isomorphism for Counterfactual Identifiability. In Singh, A.; Fazel, M.; Hsu, D.; Lacoste-Julien, S.; Berkenkamp, F.; Maharaj, T.; Wagstaff, K.; and Zhu, J., eds., *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 8022–8064. PMLR.

Cho, M.; and Sun, C. 2025. Out of Distribution Adaptation in Offline RL via Causal Normalizing Flows. *Mathematics*, 13(23).

De Sousa Ribeiro, F.; Xia, T.; Monteiro, M.; Pawlowski, N.; and Glocker, B. 2023. High Fidelity Image Counterfactuals with Probabilistic Causal Models. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 7390–7425. PMLR.

Dinh, L.; Sohl-Dickstein, J.; and Bengio, S. 2017. Density estimation using Real NVP. arXiv:1605.08803.

Fu, J.; Kumar, A.; Nachum, O.; Tucker, G.; and Levine, S. 2021. D4RL: Datasets for Deep Data-Driven Reinforcement Learning. arXiv:2004.07219.

Fujimoto, S.; and Gu, S. S. 2021. A Minimalist Approach to Offline Reinforcement Learning. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 20132–20145. Curran Associates, Inc.

Immer, A.; Schultheiss, C.; Vogt, J. E.; Schölkopf, B.; Bühlmann, P.; and Marx, A. 2023. On the Identifiability and Estimation of Causal Location-Scale Noise Models. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 14316–14332. PMLR.

Kazemi, M.; Lally, J.; Tishchenko, E.; Chockler, H.; and Paoletti, N. 2025. Counterfactual Influence in Markov Decision Processes. In Huang, B.; and Drton, M., eds., *Proceedings of the Fourth Conference on Causal Learning and Reasoning*, volume 275 of *Proceedings of Machine Learning Research*, 792–817. PMLR.

Khemakhem, I.; Monti, R.; Leech, R.; and Hyvarinen, A. 2021. Causal Autoregressive Flows. In Banerjee, A.; and Fukumizu, K., eds., *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, 3520–3528. PMLR.

Kumar, A.; Zhou, A.; Tucker, G.; and Levine, S. 2020. Conservative Q-Learning for Offline Reinforcement Learning. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1179–1191. Curran Associates, Inc.

Lally, J.; Kazemi, M.; and Paoletti, N. 2026. Robust Counterfactual Inference in Markov Decision Processes. arXiv:2502.13731.

Lee, J.; Yun, S.; Yun, T.; and Park, J. 2024. GTA: Generative Trajectory Augmentation with Guidance for Offline Reinforcement Learning. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 56766–56801. Curran Associates, Inc.

Lu, C.; Huang, B.; Wang, K.; Hernández-Lobato, J. M.; Zhang, K.; and Schölkopf, B. 2020. Sample-Efficient Reinforcement Learning via Counterfactual-Based Data Augmentation. arXiv:2012.09092.

Nasr-Esfahany, A.; Alizadeh, M.; and Shah, D. 2023. Counterfactual Identifiability of Bijective Causal Models. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 25733–25754. PMLR.

Nasr-Esfahany, A.; and Kiciman, E. 2023. Counterfactual (Non-)identifiability of Learned Structural Causal Models. arXiv:2301.09031.

Noorbakhsh, K.; and Rodriguez, M. 2022. Counterfactual Temporal Point Processes. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 24810–24823. Curran Associates, Inc.

Oberst, M.; and Sontag, D. 2019. Counterfactual Off-Policy Evaluation with Gumbel-Max Structural Causal Models. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 4881–4890. PMLR.

Pawlowski, N.; Coelho de Castro, D.; and Glocker, B. 2020. Deep Structural Causal Models for Tractable Counterfactual Inference. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 857–869. Curran Associates, Inc.

Peters, J.; Janzing, D.; and Schölkopf, B. 2017. *Elements of Causal Inference - Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning Series. Cambridge, MA, USA: The MIT Press.

Pitis, S.; Creager, E.; and Garg, A. 2020. Counterfactual Data Augmentation using Locally Factored Dynamics. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 3976–3990. Curran Associates, Inc.

Pitis, S.; Creager, E.; Mandlekar, A.; and Garg, A. 2022. MoCoDA: Model-based Counterfactual Data Augmentation. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 18143–18156. Curran Associates, Inc.

Richens, J.; Beard, R.; and Thompson, D. H. 2022. Counterfactual harm. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 36350–36365. Curran Associates, Inc.

Tsirtsis, S.; De, A.; and Rodriguez, M. 2021. Counterfactual Explanations in Sequential Decision Making Under Uncertainty. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 30127–30139. Curran Associates, Inc.

Tsirtsis, S.; and Rodriguez, M. 2023. Finding Counterfactually Optimal Action Sequences in Continuous State Spaces. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 3220–3247. Curran Associates, Inc.

Yu, T.; Kumar, A.; Rafailov, R.; Rajeswaran, A.; Levine, S.; and Finn, C. 2021. COMBO: Conservative Offline Model-Based Policy Optimization. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 28954–28967. Curran Associates, Inc.

Yu, T.; Thomas, G.; Yu, L.; Ermon, S.; Zou, J.; Levine, S.; Finn, C.; and Ma, T. 2020. MOPO: Model-based Offline Policy Optimization. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 14129–14142. Curran Associates, Inc.

Zhang, W.; SAMAD, A. M.; Xue, Y.; and Kudenko, D. 2026. Causality-Based Data-Augmentation for Offline Reinforcement Learning. *IEEE Access*, 14: 76639–76648.