Concept Modulation Models: A Unified Framework for Identifiability and Extrapolation

Soheun Yi[1]* **Yizhou Lu**[2] **Chandler Squires**[2] **Pradeep Ravikumar**[2]
[1]Department of Statistics and Data Science, Carnegie Mellon University
[2]Machine Learning Department, Carnegie Mellon University

Abstract

Reliable generalization in conditional latent variable models requires understanding both identifiability and extrapolation: how observed variation across attributes determines latent structure, and how that structure determines distributions at unseen attributes. However, existing identifiability and extrapolation guarantees are largely model-specific, with separate analyses in nonlinear ICA, causal representation learning, perturbation modeling, and related conditional latent variable models. We introduce *concept modulation models (CMMs)*, an attribute-indexed class of conditional generative models with structure $A \to \Lambda \to C \to X$, where attributes select modulators, modulators induce latent concept laws, and concepts generate observed features. CMMs lift transition-based identifiability to conditional settings by showing that feature agreement on observed attributes induces a latent concept transition constrained by the CMM class. We express these constraints through *attribute potentials*, log-density ratios between attribute-conditioned concept laws, separating the generic lifting step from model-specific rigidity arguments. The same potentials control extrapolation: agreement at unseen attributes holds exactly when the transported attribute-potential identities extend to those attributes. This yields algebraic extrapolation criteria, identifies the common potential-based proof objects behind several existing identifiability and extrapolation results, and, when combined with the model-specific rigidity arguments in those works, recovers their stated conclusions.

1 Introduction

Identifiability in latent-variable representation learning asks when latent structure is not merely useful for prediction, but uniquely determined by the variation available in data, at least up to well-characterized ambiguities [Hyvarinen et al., 2019, Khemakhem et al., 2020a, Schölkopf et al., 2021]. This question is central to reliable generalization because data from observed conditions alone need not determine behavior under unseen ones: two models may agree on all observed conditions while disagreeing off-support [D'Amour et al., 2022]. An identifiability guarantee offers a principled route around this obstacle: if the latent structure governing how attributes affect observations is identifiable, then recovering it can help justify generalization beyond the observed training conditions.

In this paper, we consider two problems central to this route in conditional generative models, where an observed feature X varies with an attribute $A \in \mathcal{A}$, such as an environment, intervention, perturbation, or conditioning input. *Identifiability* asks what aspects of the latent structure are determined, up to allowable ambiguities, by the conditional feature distributions $p(x \mid a)$ observed at attributes $a \in \mathcal{A}_o$, where $\mathcal{A}_o \subseteq \mathcal{A}$ may be a very small subset of $\mathcal{A}$. *Extrapolation* asks when agreement on the observed conditional distributions, together with the structural assumptions of the model class, forces agreement on $p(x \mid a')$ at unseen attributes $a' \in \mathcal{A} \setminus \mathcal{A}_o$.

*Correspondence: soheuny@andrew.cmu.edu.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Existing answers to these questions have largely been developed on a per-model basis. Nonlinear ICA and identifiable VAEs use auxiliary or conditioning variables to identify latent sources [Hyvarinen et al., 2019, Khemakhem et al., 2020a]. Causal representation learning uses environments or interventions to recover latent causal variables and causal structure [Ahuja et al., 2023, Squires et al., 2023, von Kügelgen et al., 2023, Varıcı et al., 2024a, 2025]. Perturbation models study when learned latent responses can extrapolate to unseen perturbations [von Kügelgen et al., 2025]. Although these results share a common reliance on structured variation across attributes, their assumptions and conclusions are usually stated in model-specific terms.

Several recent works provide unifying perspectives, but their scope is different from ours. For example, Khemakhem et al. [2020a] unify VAEs and nonlinear ICA through condition-dependent latent priors, while Yao et al. [2025] unify causal representation learning through an invariance principle. Reizinger et al. [2025] study identifiability through exchangeable mechanisms or related structural assumptions. These frameworks clarify important families of identifiability results, but they primarily address when latent representations are identifiable within particular structural regimes. Our goal is complementary: we expose the contrastive objects underlying several identifiability results and use them to derive conditions for extrapolation. We provide a more detailed survey and comparison in App. A.

To give a common formulation of these different structural regimes, we introduce a shared structural form for attribute-conditioned generation: $A \to \Lambda \to C \to X$. Here, the attribute A indexes a latent modulator Λ, the modulator specifies a latent concept distribution, and the concept C generates the observed feature X. We call such models *concept modulation models* (CMMs). The modulator separates attribute-specific indexing from the shared rule that maps modulators to concept distributions, thereby tying conditional concept laws together across attributes.

Our analysis builds on the transition-based identifiability perspective of Squires and Ravikumar [2026], which reduces latent identifiability to characterizing concept-space transitions compatible with a model class. CMMs are a conditional lift of this perspective: for each fixed attribute, a CMM induces a latent concept generative model, while the CMM class ties the family of attribute-conditioned concept laws through a shared modulation mechanism. This attribute-indexed structure raises an extrapolation question absent from the unconditional setting: when does agreement of feature distributions on observed attributes force agreement at unseen attributes?

Under the assumption that the concept laws have common support, our characterization of feature equivalence is expressed through the log-density ratio $\log p(c \mid a) - \log p(c \mid a_0)$, which we call the *attribute potential*. Attribute potentials remove terms shared across attributes, isolating how the latent concept law changes with A. We show that feature-equivalent CMMs are related by a latent transition that preserves the observed attribute potentials, so identifiability reduces to determining which transitions remain compatible with the model class.

The same object also characterizes extrapolation. Once agreement on observed attributes has been lifted to a latent transition, extrapolation is equivalent to the corresponding transported attribute-potential identities holding not only on $\mathcal{A}_o$, but also at the unseen attributes of interest. These conditions recover recent guarantees for causal representation learning and perturbation modeling, and also yield new interaction-based extrapolation guarantees for structured attribute spaces.

Our contributions are as follows:

- **Concept modulation models:** We introduce CMMs, a conditional generative framework with graphical structure $A \to \Lambda \to C \to X$ (Definitions 1 and 2). This separates attribute-specific indexing from the shared modulation mechanism and captures several weakly supervised latent-variable settings.
- **Conditional transition-based identifiability:** Building on the transition-based framework of Squires and Ravikumar [2026], we prove a conditional lifting theorem for CMMs: feature equivalence on observed attributes induces a latent concept transition compatible with both components of the model class (Thm. 1). We then characterize the concept-side compatibility condition through preservation of attribute potentials (Thm. 2).
- **Extrapolation via attribute potentials:** We prove that, once observed feature agreement has been lifted to a latent transition, agreement at unseen attributes is equivalent to extension of the transported attribute-potential identities (Thm. 3). This yields algebraic extrapola-

tion criteria, recovers perturbation extrapolation guarantees as a special case, and gives interaction-based extrapolation results for structured attribute spaces.

2 A unifying framework: Concept modulation models

2.1 Preliminaries and notation

Throughout, all spaces are standard Borel spaces. Capitals such as A, C, X denote random variables, calligraphic letters denote their spaces or classes of kernels, and bold letters such as $\mathbf{Q}, \mathbf{B}, \mathbf{K}$ denote individual kernels. For measures μ, μ' defined on a space $\mathcal{U}$, we write $\mu \ll \mu'$ if μ is absolutely continuous with respect to μ', i.e., $\mu'(B) = 0 \implies \mu(B) = 0$, and write $\mu \sim \mu'$ if $\mu \ll \mu'$ and $\mu' \ll \mu$. We say that a condition holds μ-a.e. for $u \in \mathcal{U}$ if the set of points u for which the condition fails has μ-measure zero. We write $\mu \otimes \mu'$ for the product measure of μ, μ'.

For a space $\mathcal{U}$, let $\mathcal{P}(\mathcal{U})$ denote the set of probability measures on $\mathcal{U}$. For spaces $\mathcal{U}, \mathcal{V}$, let $\mathcal{F}(\mathcal{U} \to \mathcal{V})$ denote the set of measurable maps from $\mathcal{U}$ to $\mathcal{V}$, and let $\mathcal{K}^{\mathrm{mk}}(\mathcal{U} \to \mathcal{V})$ denote the set of Markov kernels from $\mathcal{U}$ to $\mathcal{V}$. For measurable $\mathcal{U}' \subseteq \mathcal{U}$, we denote by $\mathbf{K}_{\mathcal{U}'}$ the restriction of $\mathbf{K}$ to $\mathcal{U}'$. Given a measurable map $f \in \mathcal{F}(\mathcal{U} \to \mathcal{V})$, we let $\mathbf{T}_f \colon \mathcal{P}(\mathcal{U}) \to \mathcal{P}(\mathcal{V})$ denote the associated pushforward operator, i.e., $(\mathbf{T}_f\mu)(B) := \mu(f^{-1}[B])$ for all $\mu \in \mathcal{P}(\mathcal{U})$ and measurable $B \subseteq \mathcal{V}$. Thus $\mathbf{T}_f\mu = f_{\#}\mu$ in standard pushforward notation; as a deterministic kernel, $\mathbf{T}_f(\cdot \mid u) = \delta_{f(u)}$. Given a Markov kernel $\mathbf{K} \in \mathcal{K}^{\mathrm{mk}}(\mathcal{U} \to \mathcal{V})$, we define its pushforward of measure $\mu \in \mathcal{P}(\mathcal{U})$ by $(\mathbf{K}\mu)(B) := \int_{\mathcal{U}} \mathbf{K}(B \mid u)\, \mu(du)$.

As a running example throughout this paper, we consider attribute-conditioned distributions induced by the process

$$p(c \mid a) \propto q(c) \exp\{\langle f(a), c\rangle\}, \qquad X = g(C), \tag{1}$$

where f and g are unknown. Here, the attribute a is mapped to the modulator $\lambda = f(a)$, the modulator λ changes the latent concept law through an exponential-family tilt $q(c) \exp\{\langle \cdot, c\rangle\}$, and the concept is mapped to the observation by g. In Sec. 5, we apply the framework to nonlinear ICA and iVAE, causal representation learning, and perturbation modeling.

2.2 Concept modulation models

A concept modulation model (CMM) factors a conditional generative model through $A \to \Lambda \to C \to X$, where $A \in \mathcal{A}$ is an ***attribute***, $\Lambda \in \mathscr{L}$ is a ***modulator***, $C \in \mathcal{C}$ is a latent ***concept***, and $X \in \mathcal{X}$ is an observed ***feature***. Intuitively, the attribute selects a modulator, the modulator specifies a distribution over concepts, and the concept is mapped to the feature space. Here, $\mathcal{A}$, $\mathscr{L}$, $\mathcal{C}$, and $\mathcal{X}$ are standard Borel spaces, referred to as the attribute, modulator, concept, and feature spaces, respectively. In the running example (1), the $A \to \Lambda$ stage corresponds to the map f, the $\Lambda \to C$ stage corresponds to the exponential tilting, and the $C \to X$ stage corresponds to the observation map g.

We now formalize this generative structure at the level of Markov kernels.

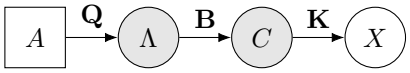
Definition 1. *A* **concept modulation model** M *on* $(\mathcal{A}, \mathscr{L}, \mathcal{C}, \mathcal{X})$ *is a tuple* $\mathsf{M} = (\mathbf{Q}, \mathbf{B}, \mathbf{K})$*, where* $\mathbf{Q} \in \mathcal{K}^{\mathrm{mk}}(\mathcal{A} \to \mathscr{L})$ *is the* **indexing kernel***,* $\mathbf{B} \in \mathcal{K}^{\mathrm{mk}}(\mathscr{L} \to \mathcal{C})$ *is the* **concept modulation kernel***, and* $\mathbf{K} \in \mathcal{K}^{\mathrm{mk}}(\mathcal{C} \to \mathcal{X})$ *is the* **mixing kernel***.*

We denote by $\mathbf{P}^{\mathsf{M}} := \mathbf{KBQ} \in \mathcal{K}^{\mathrm{mk}}(\mathcal{A} \to \mathcal{X})$ the ***feature generation kernel*** induced by M. A graphical representation of Definition 1 is shown in Figure 1: the indexing kernel $\mathbf{Q}$ specifies how attributes index modulators, while the fixed concept modulation kernel $\mathbf{B}$ specifies how each modulator induces a concept distribution. A CMM *class* specifies candidate models with a shared concept-modulation mechanism. We therefore fix the kernel $\mathbf{B}$, which maps modulators to concept distributions, while allowing the indexing kernel $\mathbf{Q}$ and the mixing kernel $\mathbf{K}$ to vary.

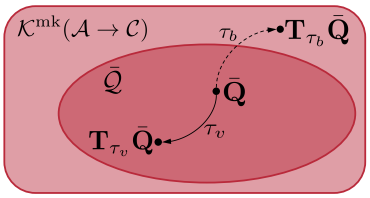
Definition 2. *A* **concept modulation model class (CMM class)** *is a product class*

$$\mathcal{M} = \mathcal{Q} \times \{\mathbf{B}\} \times \mathcal{K}, \quad \textit{where} \quad \mathcal{Q} \subseteq \mathcal{K}^{\mathrm{mk}}(\mathcal{A} \to \mathscr{L}) \quad \textit{and} \quad \mathcal{K} \subseteq \mathcal{K}^{\mathrm{mk}}(\mathcal{C} \to \mathcal{X})$$

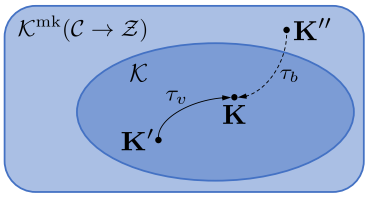
so that all models in $\mathcal{M}$ *share the same concept modulation kernel* $\mathbf{B}$*. Two models* $\mathsf{M} = (\mathbf{Q}, \mathbf{B}, \mathbf{K})$ *and* $\mathsf{M}' = (\mathbf{Q}', \mathbf{B}, \mathbf{K}')$ *in the CMM class* $\mathcal{M}$ *are* **feature equivalent on** $\mathcal{A}_{\mathrm{o}} \subseteq \mathcal{A}$ *if* $\mathbf{P}^{\mathsf{M}}_{\mathcal{A}_{\mathrm{o}}} = \mathbf{P}^{\mathsf{M}'}_{\mathcal{A}_{\mathrm{o}}}$*, equivalently, if* $\mathbf{KBQ}(\cdot \mid a) = \mathbf{K}'\mathbf{B}\mathbf{Q}'(\cdot \mid a)$ *for all* $a \in \mathcal{A}_{\mathrm{o}}$*.*



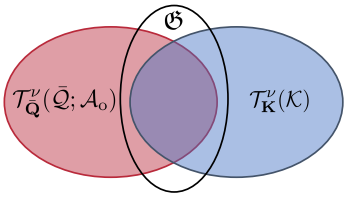
(a) Graphical model of CMM



(b) Valid and invalid concept transitions τ_v and τ_b



(c) Valid and invalid mixing transitions τ_v and τ_b



(d) Transition-intersection criterion (Thm. 1)

Figure 1: **Overview of CMMs and transition constraints.** *(a)* The observed attribute A influences a modulation variable Λ, which modulates the latent concept C, which generates the observation X. *(b,c)* Valid concept and mixing transitions are concept-space transformations that can be absorbed into the corresponding model class. Arrows in (b) denote equivalence after restriction to $\mathcal{A}_{\mathrm{o}}$, while arrows in (c) read $\mathbf{K} \overset{\nu}{=} \mathbf{K}'\mathbf{T}_{\tau_v} \overset{\nu}{=} \mathbf{K}''\mathbf{T}_{\tau_b}$. *(d)* The transition-intersection criterion selects transitions valid on both sides; if their intersection is contained in $\mathfrak{G}$, the CMM is identifiable up to $\sim_{\mathfrak{G}}$ (Thm. 1).

Given $\mathcal{M} = \mathcal{Q} \times \{\mathbf{B}\} \times \mathcal{K}$, we denote the induced concept-kernel class by $\mathbf{B}\mathcal{Q} := \{\mathbf{B}\mathbf{Q} \mid \mathbf{Q} \in \mathcal{Q}\} \subseteq \mathcal{K}^{\mathrm{mk}}(\mathcal{A} \to \mathcal{C})$. When $\mathbf{B}$ is clear from context, we write $\bar{\mathbf{Q}}$ and $\bar{\mathcal{Q}}$ for $\mathbf{B}\mathbf{Q}$ and $\mathbf{B}\mathcal{Q}$, respectively; throughout, any notation $\bar{\mathbf{Q}}$ or $\bar{\mathcal{Q}}$ is understood relative to this fixed $\mathbf{B}$.

The factorization specifies how the induced class is constructed: in (1), $\mathbf{Q}$ selects the natural parameter while the fixed $\mathbf{B}$ converts it into a normalized exponential-family law. Once $\bar{\mathcal{Q}}$ is specified, our theorems depend on $(\mathbf{B}, \mathbf{Q})$ only through $\bar{\mathbf{Q}}$ and could be stated directly for $A \to C \to X$.

As we will see, many existing identifiability results, including nonlinear ICA [Hyvarinen et al., 2019, Khemakhem et al., 2020a,b], causal representation learning [Squires et al., 2023, Buchholz et al., 2023, von Kügelgen et al., 2023, Varıcı et al., 2025], and other works [von Kügelgen et al., 2025, Schmidt et al., 2025] can be expressed in this class-level framework. Tables 2 and 3 in App. F illustrate corresponding variable-level and operator-level dictionaries for selected results.

We close this section by showing how the running example fits into the CMM framework.

Example 1. *For the running example in* (1)*, take* $\mathcal{A} = \mathbb{R}^m$*,* $\mathscr{L} = \mathbb{R}^k$*,* $\mathcal{C} = \mathbb{R}^k$*, and* $\mathcal{X} = \mathbb{R}^d$*. Then*

$$\mathbf{Q} = \mathbf{T}_f, \qquad \mathbf{B}(dc \mid \lambda) = \frac{q(c)\exp(\langle \lambda, c \rangle)}{Z(\lambda)}\, dc, \qquad \mathbf{K} = \mathbf{T}_g,$$

where $q(c) > 0$ *for all* $c \in \mathcal{C}$ *and* $Z(\lambda) = \int q(c)\exp(\langle \lambda, c \rangle)\, dc < \infty$*. A corresponding CMM class is* $\mathcal{M} = \mathcal{Q} \times \{\mathbf{B}\} \times \mathcal{K}$*, with* $\mathcal{Q} = \{\mathbf{T}_f \mid f \in \mathcal{F}\}$ *and* $\mathcal{K} = \{\mathbf{T}_g \mid g \in \mathcal{G}\}$*, where* $\mathcal{F} \subseteq \mathcal{F}(\mathcal{A} \to \mathscr{L})$ *and* $\mathcal{G} \subseteq \mathcal{F}(\mathcal{C} \to \mathcal{X})$ *are classes of functions, with* $\mathcal{G}$ *restricted to injective smooth maps.*

3 Conditional transition-based identifiability and attribute potentials

Given a CMM class, we now show how identifiability can be recovered through a common transition-intersection argument, generalizing Squires and Ravikumar [2026] to the conditional generative setting. The key idea is to view feature equivalence as evidence for a latent transition τ on concept space. For CMMs, the transition must not take the model outside the model class: after adjusting for τ, the resulting mixing kernel and attribute-conditioned concept laws must still be realizable in the model class. Identifiability up to a transition group $\mathfrak{G}$ then follows when every transition satisfying these compatibility requirements lies in $\mathfrak{G}$, as formalized in Thm. 1. The compatibility condition for the attribute-conditioned concept laws can then be characterized by log-density ratios, as in Thm. 2.

3.1 A transition-intersection criterion

We first fix the reference-measure setting in which latent transitions are defined. Let ν be a σ-finite reference measure on $\mathcal{C}$, and restrict attention to concept laws dominated by ν. For kernels $\mathbf{K}, \mathbf{K}' \in \mathcal{K}^{\mathrm{mk}}(\mathcal{C} \to \mathcal{X})$, write $\mathbf{K} \overset{\nu}{=} \mathbf{K}'$ if $\mathbf{K}\mu = \mathbf{K}'\mu$ for all $\mu \ll \nu$. A ***latent concept transition*** is a measurable automorphism of concept space, defined modulo ν-null sets:

$$\mathrm{Aut}_\nu(\mathcal{C}) := \{\tau \colon \mathcal{C} \to \mathcal{C} \mid \tau \text{ is invertible up to } \nu\text{-null sets, } \mathbf{T}_\tau \nu \sim \nu, \text{ and } \mathbf{T}_{\tau^{-1}} \nu \sim \nu\}.$$

Thus, elements of $\mathrm{Aut}_\nu(\mathcal{C})$ are precisely the concept-space transformations that preserve the ν-null sets in both directions. For example, when $\mathcal{C} = \mathbb{R}^d$ and ν is Lebesgue measure, this class contains the usual diffeomorphisms.

We next impose the structural condition that allows feature equivalence to be lifted to such a transition. Following the transition-based perspective of Squires and Ravikumar [2026], we require the mixing class to factor through an injective embedding of concept distributions.

Definition 3 (ν-Blackwell reducible mixing class)**.** *A mixing class $\mathcal{K} \subseteq \mathcal{K}^{\mathrm{mk}}(\mathcal{C} \to \mathcal{X})$ is* **ν-Blackwell reducible** *if there exist a standard Borel space $\widetilde{\mathcal{X}}$, a class $\mathcal{G}$ of bimeasurable embeddings $g \colon \mathcal{C} \to \widetilde{\mathcal{X}}$, and a shared kernel $\widetilde{\mathbf{K}} \in \mathcal{K}^{\mathrm{mk}}(\widetilde{\mathcal{X}} \to \mathcal{X})$ such that every $\mathbf{K} \in \mathcal{K}$ admits some $g \in \mathcal{G}$ with $\mathbf{K} \overset{\nu}{=} \widetilde{\mathbf{K}}\mathbf{T}_g$, and the map $\xi \mapsto \widetilde{\mathbf{K}}\xi$ is injective on $\{\mathbf{T}_g\mu \mid g \in \mathcal{G},\ \mu \ll \nu\}$.*

Injectivity of $\widetilde{\mathbf{K}}$ ensures that equality of feature laws implies equality of the corresponding pushed-forward concept laws $\mathbf{T}_g\mu$. This condition also accommodates noisy observations of the form $X = g(C) + \varepsilon$, where g is injective and $\varepsilon \sim \mathcal{N}(0, \Sigma)$ is independent of C, with $\Sigma \succ 0$ shared across candidate models. Here, $\widetilde{\mathbf{K}}$ is the Gaussian convolution channel. Since the Gaussian characteristic function is nowhere zero, $\xi * \mathcal{N}(0, \Sigma) = \xi' * \mathcal{N}(0, \Sigma)$ implies $\xi = \xi'$. Thus Blackwell reducibility holds directly for these models, including iVAE models with shared Gaussian observation noise, without requiring a deconvolution estimator.

Intuitively, a transition is valid if its effect can be absorbed without leaving the model class. On the mixing side, this means changing only the mixing kernel, so that $\mathbf{K}' \overset{\nu}{=} \mathbf{K}\mathbf{T}_\tau$ for some $\mathbf{K}' \in \mathcal{K}$. On the concept side, this means changing only the induced concept kernel, so that $\bar{\mathbf{Q}}(\cdot \mid a) = \mathbf{T}_\tau \bar{\mathbf{Q}}'(\cdot \mid a)$ for all $a \in \mathcal{A}_\mathrm{o}$ and some $\bar{\mathbf{Q}}' \in \bar{\mathcal{Q}}$.

Definition 4 (Valid transition sets)**.** *For $\bar{\mathbf{Q}} \in \bar{\mathcal{Q}}$, define the* **valid concept transitions** *by*

$$\mathcal{T}^\nu_{\bar{\mathbf{Q}}}(\bar{\mathcal{Q}}; \mathcal{A}_\mathrm{o}) := \{\tau \in \mathrm{Aut}_\nu(\mathcal{C}) \mid \exists \bar{\mathbf{Q}}' \in \bar{\mathcal{Q}} \text{ such that } \bar{\mathbf{Q}}(\cdot \mid a) = \mathbf{T}_\tau \bar{\mathbf{Q}}'(\cdot \mid a) \text{ for all } a \in \mathcal{A}_\mathrm{o}\}.$$

For $\mathbf{K} \in \mathcal{K}$, define the **valid mixing transitions** *by*

$$\mathcal{T}^\nu_{\mathbf{K}}(\mathcal{K}) := \{\tau \in \mathrm{Aut}_\nu(\mathcal{C}) \mid \exists \mathbf{K}' \in \mathcal{K} \text{ such that } \mathbf{K}' \overset{\nu}{=} \mathbf{K}\mathbf{T}_\tau\}.$$

Panels (b) and (c) of Figure 1 illustrate the valid transition sets. We adopt the notion of Blackwell equivalence proposed by Squires and Ravikumar [2026], extending the notion defined in the unconditional setup to the attribute-conditional setup.

Definition 5 (Blackwell equivalence, attribute-conditional version)**.** *When two models $\mathsf{M} = (\mathbf{Q}, \mathbf{B}, \mathbf{K})$ and $\mathsf{M}' = (\mathbf{Q}', \mathbf{B}, \mathbf{K}')$ satisfy $\mathbf{K}' \overset{\nu}{=} \mathbf{K}\mathbf{T}_\tau$ and $\bar{\mathbf{Q}}(\cdot \mid a) = \mathbf{T}_\tau \bar{\mathbf{Q}}'(\cdot \mid a)$ for all $a \in \mathcal{A}_\mathrm{o}$ with $\bar{\mathbf{Q}} = \mathbf{B}\mathbf{Q}$ and $\bar{\mathbf{Q}}' = \mathbf{B}\mathbf{Q}'$, we say M and M' are* **Blackwell equivalent via** *τ. When $\tau \in \mathfrak{G}$ for a transition group $\mathfrak{G}$, we write $\mathsf{M} \sim_{\mathfrak{G}} \mathsf{M}'$ on $\mathcal{A}_\mathrm{o}$.*

Given that $\mathfrak{G}$ is a group, $\sim_{\mathfrak{G}}$ is an equivalence relation; see App. C.1. We next impose a common-support condition to keep concept transitions and density ratios well defined. An analogous condition is assumed by Squires and Ravikumar [2026], and it holds for the main continuous examples below.

Definition 6 (Common ν-support)**.** *We say that the induced concept-kernel class $\bar{\mathcal{Q}} = \mathbf{B}\mathcal{Q}$ has* **common ν-support** *if, for every $\bar{\mathbf{Q}} \in \bar{\mathcal{Q}}$ and every $a \in \mathcal{A}$, it holds that $\bar{\mathbf{Q}}(\cdot \mid a) \ll \nu$ with density $p_{\bar{\mathbf{Q}}} := d\bar{\mathbf{Q}}(\cdot \mid a)/d\nu$ satisfying $0 < p_{\bar{\mathbf{Q}}}(c \mid a) < \infty$ for ν-almost every c.*

The following theorem is a CMM-level analogue of the transition-intersection theorem of Squires and Ravikumar [2026], visualized in panel (d) of Figure 1. When $\mathcal{A} = \mathcal{A}_\mathrm{o}$ is a singleton, the theorem reduces to the unconditional transition-intersection criterion. In the conditional setting, a single transition must transport the *entire* observed family of concept laws. Thm. 2 characterizes this transport through attribute potentials, and Thms. 3 and 4 relate the extension of these potential identities to agreement at unseen attributes.

Theorem 1 (CMM-lift of the intersection theorem)**.** *Let $\mathcal{M} = \mathcal{Q} \times \{\mathbf{B}\} \times \mathcal{K}$ be a CMM class, and let $\mathcal{A}_o \neq \varnothing$. Assume that $\mathcal{K}$ is ν-Blackwell reducible and that the induced concept-kernel class $\bar{\mathcal{Q}} = \mathbf{B}\mathcal{Q}$ has common ν-support. Then, for any $\mathsf{M} = (\mathbf{Q}, \mathbf{B}, \mathbf{K}), \mathsf{M}' = (\mathbf{Q}', \mathbf{B}, \mathbf{K}') \in \mathcal{M}$ feature-equivalent on $\mathcal{A}_o$, with $\bar{\mathbf{Q}} = \mathbf{B}\mathbf{Q}$, there exists*

$$\tau \in \mathcal{T}^{\nu}_{\mathbf{K}}(\mathcal{K}) \cap \mathcal{T}^{\nu}_{\bar{\mathbf{Q}}}(\bar{\mathcal{Q}}; \mathcal{A}_o)$$

such that M and M' are Blackwell-equivalent through τ on $\mathcal{A}_o$. If, moreover, $\mathcal{T}^{\nu}_{\mathbf{K}}(\mathcal{K}) \cap \mathcal{T}^{\nu}_{\bar{\mathbf{Q}}}(\bar{\mathcal{Q}}; \mathcal{A}_o) \subseteq \mathfrak{G}$ for a transition group $\mathfrak{G} \subseteq \mathrm{Aut}_\nu(\mathcal{C})$, then M is identifiable up to $\sim_{\mathfrak{G}}$ in $\mathcal{M}$ on $\mathcal{A}_o$, i.e., every $\mathsf{M}' \in \mathcal{M}$ feature-equivalent to M on $\mathcal{A}_o$ satisfies $\mathsf{M} \sim_{\mathfrak{G}} \mathsf{M}'$.

Proof sketch. We provide a proof sketch; App. C.2 gives the full proof under a condition weaker than the one introduced in Definition 6. Let $\mathsf{M} = (\mathbf{Q}, \mathbf{B}, \mathbf{K})$ and $\mathsf{M}' = (\mathbf{Q}', \mathbf{B}, \mathbf{K}')$ be feature-equivalent on $\mathcal{A}_o$. By ν-Blackwell reducibility, write $\mathbf{K} \overset{\nu}{=} \widetilde{\mathbf{K}}\mathbf{T}_g$ and $\mathbf{K}' \overset{\nu}{=} \widetilde{\mathbf{K}}\mathbf{T}_{g'}$, where g and g' are bimeasurable injective maps and $\widetilde{\mathbf{K}}$ is injective on the relevant pushed-forward laws. The common-support condition then makes $\tau = g^{-1} \circ g' \in \mathrm{Aut}_\nu(\mathcal{C})$ a well-defined concept transition modulo ν-null sets. We therefore obtain $\mathbf{K}' \overset{\nu}{=} \mathbf{K}\mathbf{T}_\tau$ and $\mathbf{B}\mathbf{Q}(\cdot \mid a) = \mathbf{T}_\tau \mathbf{B}\mathbf{Q}'(\cdot \mid a)$ for all $a \in \mathcal{A}_o$. Hence $\tau \in \mathcal{T}^{\nu}_{\mathbf{K}}(\mathcal{K}) \cap \mathcal{T}^{\nu}_{\bar{\mathbf{Q}}}(\bar{\mathcal{Q}}; \mathcal{A}_o)$, and if this intersection lies in $\mathfrak{G}$, the ambiguity is up to $\sim_{\mathfrak{G}}$. □

3.2 Characterizing concept-side transitions

We now characterize the concept-side valid transitions at the density level. This refines the transition-intersection criterion by isolating the common proof object behind model-specific identifiability arguments. Instead of working directly with equality of pushed concept laws, we use density ratios between attribute-conditioned concept distributions; their logarithms are the *attribute potentials*.

Definition 7. *Suppose that the induced class $\bar{\mathcal{Q}} = \mathbf{B}\mathcal{Q}$ has common ν-support as in Definition 6, and let $\mathbf{Q} \in \mathcal{Q}$. For $a, a' \in \mathcal{A}$, define the* **attribute potential** *of $\bar{\mathbf{Q}} = \mathbf{B}\mathbf{Q}$ by $\Delta^{\bar{\mathbf{Q}}}_{a,a'}(c) := \log p_{\bar{\mathbf{Q}}}(c \mid a) - \log p_{\bar{\mathbf{Q}}}(c \mid a')$, which is well-defined ν-a.e.*

The following theorem gives an equivalent characterization of the concept-side valid transition set $\mathcal{T}^{\nu}_{\bar{\mathbf{Q}}}(\bar{\mathcal{Q}}; \mathcal{A}_o)$. This characterization will also be the main tool for extrapolation in Sec. 4, where we ask when preservation of observed attribute potentials forces preservation at unseen attributes.

Theorem 2 (Density characterization of concept transitions)**.** *Consider a fixed concept modulation kernel $\mathbf{B}$, and assume that $\bar{\mathcal{Q}} = \mathbf{B}\mathcal{Q}$ has common ν-support as defined in Definition 6. Let $\mathbf{Q}, \mathbf{Q}' \in \mathcal{Q}$ be two indexing kernels. Let $\mathcal{A}' \subseteq \mathcal{A}$, $\tau \in \mathrm{Aut}_\nu(\mathcal{C})$, and define $r_{\tau^{-1}} := d(\mathbf{T}_{\tau^{-1}}\nu)/d\nu$. For any fixed anchor $a_0 \in \mathcal{A}'$ and $\bar{\mathbf{Q}} = \mathbf{B}\mathbf{Q}$, $\bar{\mathbf{Q}}' = \mathbf{B}\mathbf{Q}'$, the following are equivalent:*

(i) $\bar{\mathbf{Q}}_{\mathcal{A}'} = \mathbf{T}_\tau \bar{\mathbf{Q}}'_{\mathcal{A}'}$.

(ii) For every $a \in \mathcal{A}'$, $p_{\bar{\mathbf{Q}}'}(c \mid a) = p_{\bar{\mathbf{Q}}}(\tau(c) \mid a) r_{\tau^{-1}}(c)$ holds for ν-a.e. c.

(iii) The anchored density identity $p_{\bar{\mathbf{Q}}'}(c \mid a_0) = p_{\bar{\mathbf{Q}}}(\tau(c) \mid a_0) r_{\tau^{-1}}(c)$ holds for ν-a.e. c, and for every $a \in \mathcal{A}'$, $\Delta^{\bar{\mathbf{Q}}'}_{a,a_0}(c) = \Delta^{\bar{\mathbf{Q}}}_{a,a_0}(\tau(c))$ holds ν-a.e. c.

Consequently, $\tau \in \mathcal{T}^{\nu}_{\bar{\mathbf{Q}}}(\bar{\mathcal{Q}}; \mathcal{A}_o)$ if and only if there exists $\mathbf{Q}' \in \mathcal{Q}$ such that these equivalent conditions hold with $\mathcal{A}' = \mathcal{A}_o$.

Proof sketch. The equivalence between (i) and (ii) is the Radon–Nikodym chain rule applied to $\mathbf{T}_{\tau^{-1}}$. Under the positivity assumption, (ii) implies (iii) by taking $a = a_0$, then taking logs of the identities in (ii) for a, a_0 and subtracting to cancel $r_{\tau^{-1}}(c)$. Conversely, if (iii) holds, multiplying the anchored density identity by the exponential of the attribute potential gives (ii). The final statement follows from the definition of $\mathcal{T}^{\nu}_{\bar{\mathbf{Q}}}(\bar{\mathcal{Q}}; \mathcal{A}_o)$. □

Notably, statement *(iii)* of Thm. 2 characterizes the concept-side valid transition set $\mathcal{T}^{\nu}_{\bar{\mathbf{Q}}}(\bar{\mathcal{Q}}; \mathcal{A}_o)$ in terms of anchored density identities and attribute-potential preservation, avoiding the intractable $r_{\tau^{-1}}$ term. We now illustrate how Thms. 1 and 2 recover an identifiability result for the running example.

Example 2 (Running example, affine recovery)**.** *Consider the running example (Ex. 1). Taking* $\widetilde{\mathcal{X}} = \mathcal{X}$ *and* $\widetilde{\mathbf{K}} = \mathbf{T}_{\mathrm{id}_{\mathcal{X}}}$*, the class* $\mathcal{K}$ *is* ν*-Blackwell reducible with* ν *taken to be Lebesgue measure on* $\mathbb{R}^k$*.*

Fix $\mathsf{M} = (\mathbf{Q}, \mathbf{B}, \mathbf{K}) \in \mathcal{M}$*, with* $\mathbf{Q} = \mathbf{T}_f$ *and* $\mathbf{K} = \mathbf{T}_g$*. Fixing* $a_0 \in \mathcal{A}_{\mathrm{o}}$*, the attribute potential is*

$$\Delta^{\bar{\mathbf{Q}}}_{a,a_0}(c) = \langle f(a) - f(a_0), c\rangle - \{\log Z(f(a)) - \log Z(f(a_0))\}, \tag{2}$$

and Thm. 2 implies that for any $\tau \in \mathcal{T}^{\nu}_{\bar{\mathbf{Q}}}(\bar{\mathcal{Q}}; \mathcal{A}_{\mathrm{o}})$ *there must exist* $\mathbf{Q}' \in \mathcal{Q}$ *such that* $\Delta^{\bar{\mathbf{Q}}'}_{a,a_0}(c) = \Delta^{\bar{\mathbf{Q}}}_{a,a_0}(\tau(c))$*. In particular, if* $\mathrm{span}\{f(a) - f(a_0) \mid a \in \mathcal{A}_{\mathrm{o}}\} = \mathbb{R}^k$*, then preservation forces* τ *to be affine, as we show in App. D.1. Hence* $\mathcal{T}^{\nu}_{\bar{\mathbf{Q}}}(\bar{\mathcal{Q}}; \mathcal{A}_{\mathrm{o}}) \subseteq \mathrm{Affine}(k) \cap \mathrm{Aut}_{\nu}(\mathbb{R}^k)$*, where* $\mathrm{Affine}(k)$ *is the set of affine transformations on* $\mathbb{R}^k$*. Therefore,*

$$\mathcal{T}^{\nu}_{\mathbf{K}}(\mathcal{K}) \cap \mathcal{T}^{\nu}_{\bar{\mathbf{Q}}}(\bar{\mathcal{Q}}; \mathcal{A}_{\mathrm{o}}) \subseteq \mathcal{T}^{\nu}_{\bar{\mathbf{Q}}}(\bar{\mathcal{Q}}; \mathcal{A}_{\mathrm{o}}) \subseteq \mathfrak{G} := \mathrm{Affine}(k) \cap \mathrm{Aut}_{\nu}(\mathbb{R}^k).$$

Since the exponential-family concept laws have positive densities, the support condition in Thm. 1 holds. Therefore, Thm. 1 implies that the running example is identifiable up to $\sim_{\mathfrak{G}}$*, i.e., invertible affine transformations of the concept space.*

Many existing identifiability results impose richness assumptions under which the observed contrasts have enough rank to identify the entire latent representation up to the target ambiguity class. The attribute-potential characterization also describes what remains identifiable when this rank condition fails: the observed contrast equations constrain only the component of the latent transition visible through the observed contrast span.

Remark 1 (Partial identifiability under incomplete contrast span)**.** *Consider the CMM class from Exs. 1 and 2, but let* $V = \mathrm{span}\{f(a) - f(a_0) \mid a \in \mathcal{A}_{\mathrm{o}}\} \subsetneq \mathbb{R}^k$*. Then feature equivalence need not force the latent transition* τ *to be affine on all of* $\mathbb{R}^k$*. Nevertheless, the projection identity derived in App. D.1,* $\Pi_V \tau(c) = M \Pi_{V'} c + b$*, where* Π_V *and* $\Pi_{V'}$ *are projection matrices onto* V *and* V'*, shows that the component of the projected concept in the observed contrast subspace is still identified up to an affine map. If the observed contrast subspace is fixed across the class, i.e.,* $\mathrm{span}\{f(a) - f(a_0) \mid a \in \mathcal{A}_{\mathrm{o}}\} = V_{\mathcal{F}}$ *for every* $f \in \mathcal{F}$*, then the latent transition* τ *belongs to a group whose elements may be arbitrary off* $V_{\mathcal{F}}$ *but must act affinely after projection onto* $V_{\mathcal{F}}$*.*

4 Extrapolation via attribute potentials

Extrapolation asks when predictive agreement on observed attributes extends beyond them: if two CMMs agree on conditional feature laws over $\mathcal{A}_{\mathrm{o}}$, when must they also agree on $\mathcal{A}_{\mathrm{ex}} \supseteq \mathcal{A}_{\mathrm{o}}$? Unlike identifiability, which characterizes latent transitions compatible with feature equivalence on $\mathcal{A}_{\mathrm{o}}$, extrapolation asks whether the induced transition also enforces agreement outside $\mathcal{A}_{\mathrm{o}}$. For a transition τ from Thm. 1, with $\mathbf{K}' \overset{\nu}{=} \mathbf{K}\mathbf{T}_\tau$ and $\bar{\mathbf{Q}}_{\mathcal{A}_{\mathrm{o}}} = \mathbf{T}_\tau \bar{\mathbf{Q}}'_{\mathcal{A}_{\mathrm{o}}}$, the mixing side remains fixed as the attribute set expands. Therefore, extrapolation reduces to whether this concept-side relation extends from $\mathcal{A}_{\mathrm{o}}$ to $\mathcal{A}_{\mathrm{ex}}$. By Thm. 2, this is equivalent to preserving attribute potentials, as formalized next.

Theorem 3 (Attribute-potential characterization of extrapolation)**.** *Let* $\mathcal{M} = \mathcal{Q} \times \{\mathbf{B}\} \times \mathcal{K}$ *be a CMM class, and let* $\mathcal{A}_{\mathrm{o}} \subseteq \mathcal{A}_{\mathrm{ex}} \subseteq \mathcal{A}$*. Assume that* $\mathcal{K}$ *is* ν*-Blackwell reducible and that* $\bar{\mathcal{Q}} = \mathbf{B}\mathcal{Q}$ *has common* ν*-support. Fix* $a_0 \in \mathcal{A}_{\mathrm{o}}$*, and let* $\mathsf{M} = (\mathbf{Q}, \mathbf{B}, \mathbf{K})$ *and* $\mathsf{M}' = (\mathbf{Q}', \mathbf{B}, \mathbf{K}')$ *be feature-equivalent on* $\mathcal{A}_{\mathrm{o}}$*, and let* τ *be a transition induced by Thm. 1, so that* $\mathbf{K}' \overset{\nu}{=} \mathbf{K}\mathbf{T}_\tau$ *and* $\bar{\mathbf{Q}}_{\mathcal{A}_{\mathrm{o}}} = \mathbf{T}_\tau \bar{\mathbf{Q}}'_{\mathcal{A}_{\mathrm{o}}}$ *where* $\bar{\mathbf{Q}} = \mathbf{B}\mathbf{Q}$*,* $\bar{\mathbf{Q}}' = \mathbf{B}\mathbf{Q}'$*. Then, for this transition* τ*,*

$$\mathbf{P}^{\mathsf{M}}_{\mathcal{A}_{\mathrm{ex}}} = \mathbf{P}^{\mathsf{M}'}_{\mathcal{A}_{\mathrm{ex}}} \iff \Delta^{\bar{\mathbf{Q}}'}_{a,a_0}(c) = \Delta^{\bar{\mathbf{Q}}}_{a,a_0}(\tau(c)) \text{ for every } a \in \mathcal{A}_{\mathrm{ex}},\ \nu\text{-a.e. } c.$$

Thus, if the identities $\Delta^{\bar{\mathbf{Q}}'}_{a,a_0}(c) = \Delta^{\bar{\mathbf{Q}}}_{a,a_0}(\tau(c))$ *that hold on* $\mathcal{A}_{\mathrm{o}}$ *extend over* $\mathcal{A}_{\mathrm{ex}}$*, then* $\mathbf{P}^{\mathsf{M}}_{\mathcal{A}_{\mathrm{ex}}} = \mathbf{P}^{\mathsf{M}'}_{\mathcal{A}_{\mathrm{ex}}}$*.*

We defer the proof of Thm. 3 to App. C.4. The remaining work is to provide class-specific arguments showing that the transported attribute-potential identities on $\mathcal{A}_{\mathrm{o}}$ force the corresponding identities on $\mathcal{A}_{\mathrm{ex}}$. This implication is not automatic: in the running exponential-family CMM with unrestricted indexing map $f\colon \mathcal{A} \to \mathbb{R}^k$, one may keep $f(a)$ unchanged for all $a \in \mathcal{A}_{\mathrm{o}}$ while altering $f(a_{\mathrm{ex}})$ arbitrarily at an unseen attribute $a_{\mathrm{ex}} \notin \mathcal{A}_{\mathrm{o}}$. The observed conditional feature laws are unchanged, but $p(X \mid A = a_{\mathrm{ex}})$ can change. Thus extrapolation requires structural assumptions tying unseen attribute potentials to observed ones. The following theorem gives one such condition, requiring affine dependence through a fixed attribute representation φ.

Theorem 4 (Attribute-potential preservation under affine attribute-potential differences)**.** *Let $\mathcal{M} = \mathcal{Q} \times \{\mathbf{B}\} \times \mathcal{K}$ be a CMM class satisfying the assumptions of Thms. 1 and 2. Let $\mathcal{A}_o \subseteq \mathcal{A}_{ex} \subseteq \mathcal{A}$, and fix $a_0 \in \mathcal{A}_o$. Suppose there exists a fixed attribute representation $\varphi\colon \mathcal{A} \to \mathbb{R}^m$ such that, for every $\bar{\mathbf{Q}} \in \bar{\mathcal{Q}} = \mathbf{B}\mathcal{Q}$, there exists a measurable map $D^{\bar{\mathbf{Q}}}\colon \mathcal{C} \times \mathcal{C} \to \mathbb{R}^m$ satisfying*

$$\Delta^{\bar{\mathbf{Q}}}_{a,a_0}(c) - \Delta^{\bar{\mathbf{Q}}}_{a,a_0}(c_0) = \langle \varphi(a) - \varphi(a_0), D^{\bar{\mathbf{Q}}}(c, c_0)\rangle$$

for every $a \in \mathcal{A}_{ex}$ and $\nu \otimes \nu$-a.e. (c, c_0). Let $\mathsf{M} = (\mathbf{Q}, \mathbf{B}, \mathbf{K})$ and $\mathsf{M}' = (\mathbf{Q}', \mathbf{B}, \mathbf{K}')$ be feature-equivalent on $\mathcal{A}_o$, and let τ be the transition induced by Thm. 1. If $\varphi(\mathcal{A}_{ex})$ is contained in the affine hull of $\varphi(\mathcal{A}_o)$, then the transported attribute-potential identities extend from $\mathcal{A}_o$ to $\mathcal{A}_{ex}$: for every $a \in \mathcal{A}_{ex}$, $\Delta^{\bar{\mathbf{Q}}'}_{a,a_0}(c) = \Delta^{\bar{\mathbf{Q}}}_{a,a_0}(\tau(c))$ for ν-a.e. c.

Proof sketch. We defer the full proof to App. C.5. For any $a_{ex} \in \mathcal{A}_{ex}$, we can write $\varphi(a_{ex}) = \sum_i \alpha_i \varphi(a_i)$, where $a_i \in \mathcal{A}_o$ and $\sum_i \alpha_i = 1$. For any $\mathsf{M} \in \mathcal{M}$, we have the centered potential $\Delta^{\bar{\mathbf{Q}}}_{a,a_0}(c) - \Delta^{\bar{\mathbf{Q}}}_{a,a_0}(c_0) = \langle \varphi(a) - \varphi(a_0), D^{\bar{\mathbf{Q}}}(c, c_0)\rangle$, so the same quantity at a_{ex} is an affine combination of the centered potentials at the a_i's, and the same relation holds for $\bar{\mathbf{Q}}'$. Statement *(iii)* of Thm. 2 transports the observed identities, hence the affine expansion gives $\Delta^{\bar{\mathbf{Q}}'}_{a_{ex},a_0}(c) - \Delta^{\bar{\mathbf{Q}}'}_{a_{ex},a_0}(c_0) = \Delta^{\bar{\mathbf{Q}}}_{a_{ex},a_0}(\tau(c)) - \Delta^{\bar{\mathbf{Q}}}_{a_{ex},a_0}(\tau(c_0))$. The density identity at a_0 and normalization at a_{ex} together force $\Delta^{\bar{\mathbf{Q}}'}_{a_{ex},a_0}(c) = \Delta^{\bar{\mathbf{Q}}}_{a_{ex},a_0}(\tau(c))$. □

The point of Thm. 4 is that extrapolation is driven by the linear structure of attribute potentials, not by linearity of the raw attribute space. The representation φ may encode nonlinear or combinatorial structure, such as interaction terms among atomic attributes, as in factorial designs [Fisher, 1935] and combinatorial intervention models [O'Donnell, 2008, Agarwal et al., 2023]. Thus, the theorem lifts feature-equivalence on finitely many interactions to extrapolation guarantees over a much larger combinatorial attribute space, as stated in the following corollary.

Corollary 1 (Interaction-basis extrapolation)**.** *Let $\mathcal{M} = \mathcal{Q} \times \{\mathbf{B}\} \times \mathcal{K}$ satisfy the assumptions of Thm. 3. Let $\mathcal{A} = 2^{[m]}$, let $\mathcal{H} \subseteq 2^{[m]}$ be a finite downward-closed family containing $\varnothing$, i.e., $H \in \mathcal{H}$ and $U \subseteq H$ imply $U \in \mathcal{H}$, and set $\mathcal{A}_o = \mathcal{H}$. Assume every $\mathsf{M} \in \mathcal{M}$ induces concept densities of the form $p^{\mathsf{M}}(c \mid S) = \exp\left(\sum_{T \in \mathcal{H},\, T \subseteq S} h^{\mathsf{M}}_T(c)\right) / Z^{\mathsf{M}}(S)$. Then feature equivalence on $\mathcal{A}_o$ implies feature equivalence on all $S \subseteq [m]$.*

Proof sketch. We defer a full proof to App. C.6. Denoting $\mathbf{1}$ by the indicator function, we have $\Delta^{\mathsf{M}}_{S,\varnothing}(c) - \Delta^{\mathsf{M}}_{S,\varnothing}(c_0) = \langle \varphi_{\mathcal{H}}(S) - \varphi_{\mathcal{H}}(\varnothing), D^{\mathsf{M}}(c, c_0)\rangle$, where $\varphi_{\mathcal{H}}(S) = (\mathbf{1}\{T \subseteq S\})_{T \in \mathcal{H}}$ and $D^{\mathsf{M}}_T(c, c_0) = h^{\mathsf{M}}_T(c) - h^{\mathsf{M}}_T(c_0)$ for $T \neq \varnothing$ and $D^{\mathsf{M}}_{\varnothing}(c, c_0) = 0$. One can show that $\varphi_{\mathcal{H}}(S)$ is an affine combination of $\{\varphi_{\mathcal{H}}(U) \mid U \in \mathcal{H}\}$ for any $S \subseteq [m]$. Thms. 3 and 4 yield desired result. □

We close this section by recording an implication of Thm. 4 for the running example.

Example 3 (Running example: extrapolation)**.** *For the running exponential-family CMM, assume that $f(a) = W_f \varphi(a)$ for a matrix W_f depending on f and a fixed attribute representation φ. This satisfies the attribute-potential assumption of Thm. 4 with $D^{\bar{\mathbf{Q}}}(c, c_0) = W_f^\top (c - c_0)$. Consequently, by Thm. 3, feature equivalence on $\mathcal{A}_o$ implies feature equivalence at every target $a_{ex} \in \mathcal{A}_{ex}$ whose $\varphi(a_{ex})$ is an affine combination of $\{\varphi(a) \mid a \in \mathcal{A}_o\}$.*

5 Applications: Recoveries and examples

In this section, we apply the transition-and-potential characterizations from Thms. 1 and 2 and the extrapolation criteria from Thms. 3 and 4 to representative model classes. Identifiability follows by combining the induced attribute potentials with class-specific rigidity arguments, typically drawn from the cited works. Extrapolation follows by checking whether the transported potential identities observed on $\mathcal{A}_o$ force the corresponding identities on target attributes in $\mathcal{A}_{ex}$.

Nonlinear ICA and iVAE. Nonlinear ICA with auxiliary variables [Hyvarinen et al., 2019] and iVAE [Khemakhem et al., 2020a] fit the CMM template by taking the attribute $a \in \mathcal{A}$ to be the auxiliary or conditioning variable, the concept $c = (c_1, \dots, c_k) \in \mathcal{C}$ to be the latent source vector, and the feature $x \in \mathcal{X}$ to be the nonlinear observation. In the conditionally independent source setting, the attribute potential decomposes as $\Delta^{\bar{\mathbf{Q}}}_{a,a_0}(c) = \sum_{i=1}^{k} [\log q_i(c_i \mid a) - \log q_i(c_i \mid a_0)]$, so the CMM contrast is exactly the coordinate-wise auxiliary-variable contrast used in nonlinear ICA and iVAE. Under the variability assumptions of Hyvarinen et al. [2019], preservation of these potentials rules out mixing across source coordinates, forcing $\tau(c) = (\phi_1(c_{\pi(1)}), \dots, \phi_k(c_{\pi(k)}))$ up to a permutation π and invertible scalar maps ϕ_i. More details on recoveries are given in App. E.1.

Causal representation learning. Causal representation learning fits the CMM template by taking $a \in \mathcal{A}$ to be an environment or intervention label, $c = (c_1, \dots, c_k) \in \prod_{i=1}^{k} \mathcal{C}_i$ to be the latent causal representation, and $x \in \mathcal{X}$ to be the observed representation. We work in the common-support regime, so the induced joint concept laws have positive densities with respect to a common reference measure and the attribute potentials in Thm. 2 are well defined. The modulator is the tuple of local mechanisms $\lambda = (p_i)_{i=1}^{k}$, indexed by a directed acyclic graph (DAG) G_λ, with $\mathscr{L} = \bigsqcup_{G \in \mathrm{DAG}([k])} \prod_{i=1}^{k} \mathcal{K}^{\mathrm{mk}}(\mathcal{C}_{\mathrm{pa}_G(i)} \to \mathcal{C}_i)$ and $\mathbf{B}(dc \mid \lambda) = \prod_{i=1}^{k} p_i(c_i \mid c_{\mathrm{pa}_{G_\lambda}(i)})\, dc$. Here, $\mathrm{DAG}([k])$ is the set of DAGs on $[k]$, $\mathrm{pa}_G(i)$ is the parent set of node i in G, $\mathcal{C}_S := \prod_{i \in S} \mathcal{C}_i$, and the indexing kernel $\mathbf{Q} = \mathbf{T}_f$ maps each environment label to its active tuple of local mechanisms.

For an intervention label a, let $\mathrm{tar}(a) \subseteq [k]$ denote its target nodes, and write p_i^a for the local mechanism active at node i under a: it satisfies $p_i^a = p_i^0$ for $i \notin \mathrm{tar}(a)$, while p_i^a may differ from p_i^0 for $i \in \mathrm{tar}(a)$. Fixing an observational label $a_0 \in \mathcal{A}_\mathrm{o}$ with mechanisms $(p_i^0)_{i=1}^{k}$, the attribute potential becomes $\Delta^{\bar{\mathbf{Q}}}_{a,a_0}(c) = \sum_{i \in \mathrm{tar}(a)} [\log p_i^a(c_i \mid c_{\mathrm{pa}_G(i)}) - \log p_i^0(c_i \mid c_{\mathrm{pa}_G(i)})]$. Thus, the attribute potential records exactly the local mechanism contrasts induced by the intervention. Across CRL variants, identifiability follows from showing that transitions preserving these contrasts must belong to the target ambiguity class: Gaussian CRL [Buchholz et al., 2023], interventional-to-observational density ratios in nonparametric CRL [von Kügelgen et al., 2023], and gradients of environment ratios in score-based CRL [Varıcı et al., 2024b, 2025].

For extrapolation, the additional structure is supplied by the independent causal mechanisms (ICM) principle [Peters et al., 2017, Schölkopf et al., 2021]: interventions may replace selected local mechanisms while leaving the remaining mechanisms unchanged. Let $J \subseteq [k]$ be a set of intervention labels such that $\{a_j \mid j \in J\} \subseteq \mathcal{A}_\mathrm{o}$ have pairwise disjoint targets, i.e., $\mathrm{tar}(a_j) \cap \mathrm{tar}(a_{j'}) = \varnothing$ for all $j \neq j'$. The associated composite intervention a_J is defined by $p_i^{a_J} = p_i^{a_j}$ if $i \in \mathrm{tar}(a_j)$ for the unique $j \in J$, and $p_i^{a_J} = p_i^0$ if $i \notin \bigcup_{j \in J} \mathrm{tar}(a_j)$. Under the ICM principle, non-target mechanisms cancel against the anchor and the target sets do not overlap, so the composite potential is additive: $\Delta^{\bar{\mathbf{Q}}}_{a_J,a_0}(c) = \sum_{j \in J} \Delta^{\bar{\mathbf{Q}}}_{a_j,a_0}(c)$. Therefore, if the transported identities $\Delta^{\bar{\mathbf{Q}}'}_{a_j,a_0}(c) = \Delta^{\bar{\mathbf{Q}}}_{a_j,a_0}(\tau(c))$ hold for all $j \in J$, additivity gives the same identity for a_J. By Thm. 3, feature equivalence on the observed environments then extrapolates to the composite intervention a_J. We note that this is analogous to the intervention-generalization strategy of Bravo-Hermsdorff et al. [2023], which studies extrapolation under factorization assumptions on the interventional factor models.

Perturbation modeling. von Kügelgen et al. [2025] model perturbations as shifts in the mean of a Gaussian latent distribution. This fits the CMM template by taking $a \in \mathbb{R}^K$ to encode a perturbation, $C \in \mathbb{R}^k$ to be the latent state, and $X = g(C)$ to be the observed measurement, with injective g. The indexing kernel is $\mathbf{Q} = \mathbf{T}_{a \mapsto W(a - a_0)}$, and the shared concept-modulation kernel is $\mathbf{B}(dc \mid \lambda) = \mathcal{N}(\lambda, I_k)(dc)$. Lebesgue measure is a common reference with positive finite densities, and injective mixing verifies Blackwell reducibility. The same verification applies to $X = g(C) + \varepsilon$ with a shared independent Gaussian noise law, as above. Writing $v = a - a_0$, direct calculation gives

$$\Delta^{\bar{\mathbf{Q}}}_{a,a_0}(c) = \langle Wv, c \rangle - \tfrac{1}{2}\|Wv\|^2, \qquad \langle W'v, c \rangle - \tfrac{1}{2}\|W'v\|^2 = \langle Wv, \tau(c) \rangle - \tfrac{1}{2}\|Wv\|^2.$$

The second identity is the observed transition constraint from Thm. 2. If the observed Wv span $\mathbb{R}^k$, it forces τ to be affine; transport of the anchor law $\mathcal{N}(0, I_k)$ then forces $\tau(c) = Oc$ with orthogonal O. Such rotations are valid with $W' = O^\top W$ when the decoder class permits $g' = g \circ O$; scaling by 2 is invalid because it changes the anchor covariance. After centering at c_0, the potential is $\langle a - a_0, W^\top(c - c_0) \rangle$. Thus Thm. 4 applies with $\varphi(a) = a$ and yields extrapolation whenever $a_\mathrm{ex} - a_0 \in \mathrm{span}\{a - a_0 \mid a \in \mathcal{A}_\mathrm{o}\}$, without requiring full latent-span identifiability. Under the full

latent-span condition above, perturbation effects are identified up to orthogonal ambiguity on the observed attribute span.

Factorial designs organize interventions into combinations [Dasgupta et al., 2015], while related work studies prediction at unobserved combinations [Agarwal et al., 2023, Lotfollahi et al., 2023]. We apply the same extrapolation criterion to a CMM with only main and pairwise effects in the latent log-density. Consider three perturbations, with $S \subseteq [3]$ recording the active combination, C the latent state, and X its observed measurement. The modulator is the combined interaction energy, giving $p^{\mathsf{M}}(c \mid S) \propto q(c) \exp\{\sum_{i \in S} h_i^{\mathsf{M}}(c) + \sum_{\{i,j\} \subseteq S} h_{ij}^{\mathsf{M}}(c)\}$. Writing $\widetilde{\Delta}_S^{\mathsf{M}}(c, c_0) = \Delta_{S,\varnothing}^{\mathsf{M}}(c) - \Delta_{S,\varnothing}^{\mathsf{M}}(c_0)$ gives

$$\widetilde{\Delta}_{123}^{\mathsf{M}} = \widetilde{\Delta}_{12}^{\mathsf{M}} + \widetilde{\Delta}_{13}^{\mathsf{M}} + \widetilde{\Delta}_{23}^{\mathsf{M}} - \widetilde{\Delta}_{1}^{\mathsf{M}} - \widetilde{\Delta}_{2}^{\mathsf{M}} - \widetilde{\Delta}_{3}^{\mathsf{M}}.$$

If both candidate models obey this interaction restriction and the common-support and observation-channel assumptions, agreement at control, singleton, and pair conditions forces agreement at the held-out triple by Cor. 1, yielding extrapolation to the unseen perturbation combination. We detail this example in App. D.2.

6 Discussion

CMMs provide a common language for studying identifiability and extrapolation through attribute potentials. Their main role is to separate the generic transition-and-contrast step from model-specific rigidity arguments: feature agreement first induces a latent transition, and the remaining question is whether the model class forces this transition to preserve the relevant attribute potentials.

Our results assume exact agreement of the observed feature laws within a specified CMM class. Extending these population-level guarantees to models fitted from finite data would require bounds on how estimation error and model misspecification affect latent transitions and predictions at unseen attributes.

Our density-ratio formulation uses a common-support assumption. This covers many exponential-family, soft-intervention, and stochastic hard-intervention examples, but does not cover deterministic do-interventions or other support-changing mechanisms. Extending the theory to singular or partially overlapping concept laws would broaden its applicability to causal representation learning without changing the central transition-and-contrast viewpoint.

A further direction is to use CMMs as identifiability guidance for weakly supervised and constraint-aware representation learning. Recent neurosymbolic and concept-based systems seek to learn intermediate symbolic or concept-level representations from indirect supervision, such as labels, answers, captions, or consistency constraints [Duan et al., 2023, Daniele et al., 2023, Barbiero et al., 2023, Oikarinen et al., 2023]. These works show that symbolic or concept-level structure can guide representation learning, but their results typically concern optimization, interpretability, or empirical generalization, rather than rigorous identifiability and extrapolation guarantees. Our framework suggests one route toward such guarantees by specifying how tasks or attributes modulate latent concept laws and characterizing the induced attribute potentials.

Acknowledgments and Disclosure of Funding

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA) via HR0011-25-3-0239, FA8750-23-2-1015, ONR via N00014-23-1-2368, and NSF via IIS-1909816.

References

Emre Acartürk, Burak Varıcı, Karthikeyan Shanmugam, and Ali Tajer. Sample complexity of interventional causal representation learning. In *Advances in Neural Information Processing Systems*, pages 39350–39385, 2024.

Abhineet Agarwal, Anish Agarwal, and Suhas Vijaykumar. Synthetic combinations: A causal inference framework for combinatorial interventions. *Advances in Neural Information Processing Systems*, pages 19195–19216, 2023.

Kartik Ahuja, Jason S. Hartford, and Yoshua Bengio. Properties from mechanisms: An equivariance perspective on identifiable representation learning. In *International Conference on Learning Representations*, 2022a.

Kartik Ahuja, Jason S Hartford, and Yoshua Bengio. Weakly supervised representation learning with sparse perturbations. *Advances in Neural Information Processing Systems*, pages 15516–15528, 2022b.

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International Conference on Machine Learning*, pages 372–407, 2023.

Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models – going beyond SVD. In *IEEE Annual Symposium on Foundations of Computer Science*, pages 1–10, 2012.

Pietro Barbiero, Gabriele Ciravegna, Francesco Giannini, Mateo Espinosa Zarlenga, Lucie Charlotte Magister, Alberto Tonda, Pietro Lio, Frederic Precioso, Mateja Jamnik, and Giuseppe Marra. Interpretable neural-symbolic concept reasoning. In *International Conference on Machine Learning*, pages 1801–1825, 2023.

Gecia Bravo-Hermsdorff, David Watson, Jialin Yu, Jakob Zeitler, and Ricardo Silva. Intervention generalization: A view from factor graph models. In *Advances in Neural Information Processing Systems*, pages 43662–43675, 2023.

Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*, pages 38319–38331, 2022.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Kumar Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. In *Advances in Neural Information Processing Systems*, pages 45419–45462, 2023.

Jingyi Cui, Qi Zhang, Yifei Wang, and Yisen Wang. On the limits of sparse autoencoders: A theoretical framework and reweighted remedy. In *International Conference on Learning Representations*, 2026.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Alessandro Daniele, Tommaso Campari, Sagar Malhotra, and Luciano Serafini. Deep symbolic learning: discovering symbols and rules from perceptions. In *International Joint Conference on Artificial Intelligence*, 2023.

Tirthankar Dasgupta, Natesh S. Pillai, and Donald B. Rubin. Causal inference from 2^K factorial designs by using potential outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(4):727–753, 2015.

Yilun Du and Leslie Kaelbling. Position: compositional generative modeling: a single model is not all you need. In *International Conference on Machine Learning*, 2024.

Xuguang Duan, Xin Wang, Peilin Zhao, Guangyao Shen, and Wenwu Zhu. Deeplogic: Joint learning of neural perception and logical reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4321–4334, 2023.

R. A. Fisher. *The Design of Experiments*. The Design of Experiments. Oliver & Boyd, Oxford, England, 1935.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, 2025.

Kejun Huang, Xiao Fu, and Nikolaos D Sidiropoulos. Anchor-free correlated topic modeling: Identifiability and algorithm. *Advances in Neural Information Processing Systems*, 2016.

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

Jikai Jin and Vasilis Syrgkanis. Learning linear causal representations from general environments: Identifiability and intrinsic ambiguity. In *Advances in Neural Information Processing Systems*, pages 63466–63509, 2024.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217, 2020a.

Ilyes Khemakhem, Ricardo Monti, Diederik Kingma, and Aapo Hyvarinen. ICE-BeeM: Identifiable conditional energy-based deep models based on nonlinear ICA. *Advances in Neural Information Processing Systems*, pages 12768–12778, 2020b.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning*, pages 2668–2677, 2018.

Lingjing Kong, Shaoan Xie, Weiran Yao, Yujia Zheng, Guangyi Chen, Petar Stojanov, Victor Akinwande, and Kun Zhang. Partial disentanglement for domain adaptation. In *International Conference on Machine Learning*, pages 11455–11472, 2022.

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

Sebastien Lachapelle, Divyat Mahajan, Ioannis Mitliagkas, and Simon Lacoste-Julien. Additive decoders for latent variables identification and cartesian-product extrapolation. In *Advances in Neural Information Processing Systems*, pages 25112–25150, 2023.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. CITRIS: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning*, pages 13557–13603, 2022.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. BISCUIT: Causal representation learning from binary interactions. In *Conference on Uncertainty in Artificial Intelligence*, pages 1263–1273, 2023.

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

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L. Ibarra, Sanjay R. Srivatsan, Mohsen Naghipourfar, Riza M. Daza, Beth Martin, Jay Shendure, Jose L. McFaline-Figueroa, Pierre Boyeau, F. Alexander Wolf, Nafissa Yakubova, Stephan Günnemann, Cole Trapnell, David Lopez-Paz, and Fabian J. Theis. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6):e11517, 2023.

Ryan O'Donnell. Some topics in analysis of boolean functions. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, pages 569–578. Association for Computing Machinery, 2008.

Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *International Conference on Learning Representations*, 2023.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning Series. MIT Press, Cambridge, MA, USA, 2017.

Goutham Rajendran, Simon Buchholz, Bryon Aragam, Bernhard Schölkopf, and Pradeep Kumar Ravikumar. From causal to concept-based representation learning. In *Advances in Neural Information Processing Systems*, pages 101250–101296, 2024.

Patrik Reizinger, Siyuan Guo, Ferenc Huszár, Bernhard Schölkopf, and Wieland Brendel. Identifiable exchangeable mechanisms for causal structure and representation learning. In *International Conference on Learning Representations*, 2025.

Gian-Carlo Rota. On the foundations of combinatorial theory I. Theory of Möbius Functions. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 2:340–368, 1964.

Sorawit Saengkyongam, Elan Rosenfeld, Pradeep Kumar Ravikumar, Niklas Pfister, and Jonas Peters. Identifying representations for intervention extrapolation. In *International Conference on Learning Representations*, 2024.

Tobias Schmidt, Steffen Schneider, and Matthias Bethge. Equivariance by contrast: Identifiable equivariant embeddings from unlabeled finite group actions. In *Advances in Neural Information Processing Systems*, pages 71760–71800, 2025.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Chandler Squires and Pradeep Ravikumar. A unifying framework for unsupervised concept extraction, 2026. URL https://arxiv.org/abs/2604.24936.

Chandler Squires, Anna Seigal, Salil S Bhate, and Caroline Uhler. Linear causal disentanglement via interventions. In *International Conference on Machine Learning*, pages 32540–32560, 2023.

Burak Varıcı, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. General identifiability and achievability for causal representation learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2314–2322, 2024a.

Burak Varıcı, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. Linear causal representation learning from unknown multi-node interventions. *Advances in Neural Information Processing Systems*, pages 111614–111648, 2024b.

Burak Varıcı, Emre Acartürk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based causal representation learning: Linear and general transformations. *Journal of Machine Learning Research*, 26(112):1–90, 2025.

Julius von Kügelgen, Michel Besserve, Wendong Liang, Luigi Gresele, Armin Kekić, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. In *Advances in Neural Information Processing Systems*, pages 48603–48638, 2023.

Julius von Kügelgen, Xinwei Shen, Jakob Ketterer, Nicolai Meinshausen, and Jonas Peters. Representation learning for distributional perturbation extrapolation. In *Learning Meaningful Representations of Life (LMRL) Workshop at ICLR*, 2025.

Danru Xu, Dingling Yao, Sébastien Lachapelle, Perouz Taslakian, Julius von Kügelgen, Francesco Locatello, and Sara Magliacane. A sparsity principle for partially observable causal representation learning. In *International Conference on Machine Learning*, 2024.

Dingling Yao, Danru Xu, Sébastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. In *International Conference on Learning Representations*, 2024.

Dingling Yao, Dario Rancati, Riccardo Cadei, Marco Fumero, and Francesco Locatello. Unifying causal representation learning with the invariance principle. In *International Conference on Learning Representations*, 2025.

Jiaqi Zhang, Kristjan Greenewald, Chandler Squires, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. *Advances in Neural Information Processing Systems*, pages 50254–50292, 2023.

Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting. In *International Conference on Machine Learning*, 2024.

Yujia Zheng, Shaoan Xie, and Kun Zhang. Nonparametric identification of latent concepts. In *International Conference on Machine Learning*, pages 78374–78404, 2025.

Contents of Appendix

A Related works **16**

A.1 Identifiability in latent-variable representation learning 16

A.2 Causal representation learning 16

A.3 Concept extraction 16

A.4 Extrapolation 17

A.5 Unifying frameworks 17

B Preliminaries **18**

B.1 Notation summary 18

B.2 Measure-theoretic background 18

C Proofs **19**

C.1 Proof of $\sim_{\mathfrak{G}}$ being equivalence relation 19

C.2 Proof of Thm. 1 19

C.3 Proof of Thm. 2 21

C.4 Proof of Thm. 3 22

C.5 Proof of Thm. 4 23

C.6 Proof of Cor. 1 24

D Deferred details **25**

D.1 Details for the running example 25

D.2 Details for factorial perturbation modeling 26

E Recoveries of representative prior work **27**

E.1 Nonlinear ICA 28

E.2 Conditionally exponential families 29

E.3 Linear CRL 30

E.4 Gaussian CRL 30

E.5 Score-based CRL 32

E.6 Nonparametric CRL 33

E.7 Perturbation modeling 34

F CMM Translations of prior work **35**

A Related works

A.1 Identifiability in latent-variable representation learning

Identifiability in latent-variable representation learning asks when latent structure is determined by observed data, up to an allowable ambiguity class such as permutation, component-wise transformations, or affine transformations. A canonical example is nonlinear ICA, where statistically independent latent components are observed only through a nonlinear mixing map. Without additional structure, nonlinear ICA and unsupervised disentanglement are generally unidentifiable [Locatello et al., 2019]. A central way to restore identifiability is to use auxiliary or conditioning information. Hyvarinen et al. [2019] show that an observed auxiliary variable can identify nonlinear ICA representations when latent components are conditionally independent and sufficiently modulated by the auxiliary variable. Khemakhem et al. [2020a] develop this principle in a variational latent-variable framework, using a condition-dependent exponential-family prior and an injective decoder to obtain identifiability. Khemakhem et al. [2020b] instantiate the same idea in conditional energy-based models, where structured conditioning of the energy function yields identifiable representations. A complementary view derives identifiability from restrictions on latent mechanisms and their symmetries rather than from a specific exponential-family form [Ahuja et al., 2022a]. Our framework follows this broader identifiability perspective, but studies conditional generative models indexed by attributes and asks both which latent concept structure is identifiable from observed attributes and when this structure extrapolates to unseen attributes.

A.2 Causal representation learning

Causal representation learning (CRL) seeks to recover latent causal variables, and sometimes their causal graph, from high-dimensional observations [Schölkopf et al., 2021]. Interventional CRL uses environments or interventions as the source of variation, but the broader literature also exploits temporal structure, paired intervention data, distribution shifts, multi-view observations, and partial observability. In the static interventional setting, Squires et al. [2023] study linear latent causal models under injective linear mixing and show that one perfect intervention per latent variable is sufficient, and in a worst-case sense necessary, for identifying the latent causal model and the mixing map. Ahuja et al. [2023] study interventional CRL under perfect and imperfect interventions, showing how intervention-induced geometric structure can identify latent causal variables up to appropriate ambiguities. Buchholz et al. [2023] allow general nonlinear mixing while retaining a linear-Gaussian latent causal model, and obtain identifiability from unpaired interventional data with unknown single-node intervention targets. von Kügelgen et al. [2023] move to nonparametric latent causal models and diffeomorphic mixing under unknown interventions, clarifying both positive identifiability results and residual equivalence classes. Other work relaxes or changes the intervention model, including soft interventions [Zhang et al., 2023], uncoupled hard interventions and general nonparametric models [Varıcı et al., 2024a], unknown multi-node interventions under linear transformations [Varıcı et al., 2024b], and score-based identification under linear and general nonlinear transformations [Varıcı et al., 2025]. Complementary CRL results use weaker or different supervision signals. Brehmer et al. [2022] identify latent causal variables and structure from paired pre- and post-intervention observations, generated from a shared exogenous noise, with random unknown interventions. Lippe et al. [2022] use temporal sequences with intervention targets to identify causal factors, while Lippe et al. [2023] replaces observed targets with binary interaction variables. Multi-view and multi-distribution approaches show that explicit intervention labels are not the only route to identifiability, with guarantees under partial observability [Yao et al., 2024, Xu et al., 2024] and general distribution shifts [Zhang et al., 2024]. Recent work also studies intrinsic ambiguity and finite-sample behavior in interventional CRL [Jin and Syrgkanis, 2024, Acartürk et al., 2024]. These works provide model-specific identifiability guarantees under particular sources of variation, whereas CMMs isolate a generic lifting step from observed feature agreement to constrained latent concept transitions and leave model-specific rigidity to separate assumptions.

A.3 Concept extraction

Our use of the term "concept" is related to work on extracting latent concepts or factors from observed representations, but our goal is different. Concept-extraction methods in interpretability often aim to find human-facing features in neural activations using examples, labels, sparsity, or

dictionary-learning objectives [Kim et al., 2018, Oikarinen et al., 2023, Bricken et al., 2023, Gao et al., 2025]. Recent theoretical work instead asks when such latent concepts are identifiable. Cui et al. [2026] analyze identifiability limits for sparse autoencoders, Zheng et al. [2025] give nonparametric guarantees for identifying latent concepts, and Squires and Ravikumar [2026] provide a transition-based framework for unsupervised concept extraction. Rajendran et al. [2024] connect CRL and concept-based representation learning by relaxing recovery of true causal variables toward recovery of task-relevant concepts represented in latent space. Lachapelle et al. [2022] show that mechanism sparsity can also drive disentanglement in nonlinear ICA, illustrating how sparse modulation of latent components can reduce non-identifiability. CMMs differ from these works by making the concept law attribute-indexed: attributes modulate latent concept distributions, so the central questions are not only concept identifiability from observed attributes but also extrapolation of concept and feature laws to unseen attributes.

A.4 Extrapolation

A growing line of work studies when models learned from observed conditions can predict behavior under unseen interventions, perturbations, or attribute values. Lachapelle et al. [2023] study additive decoders and show that additive latent structure can support Cartesian-product extrapolation by recombining observed factors of variation. Agarwal et al. [2023] study combinatorial interventions from a causal-inference perspective, using low-rank and Fourier-sparsity structure to identify potential outcomes for unseen intervention combinations. Du and Kaelbling [2024] discuss compositional generative modeling more broadly, emphasizing how modular structure can support generalization to unseen combinations. In latent-variable settings, Saengkyongam et al. [2024] establish affine identifiability of latent representations under linear intervention effects and use this to certify extrapolation to out-of-support intervention values for downstream outcomes. von Kügelgen et al. [2025] study distributional perturbation extrapolation under latent mean-shift models and prove representation-identifiability and perturbation-extrapolation guarantees. Kong et al. [2022] study domain adaptation under partial disentanglement, where target-domain observations are available and the goal is prediction rather than identification of a full unseen interventional distribution. These works give extrapolation guarantees in specific structural regimes. CMMs instead formulate extrapolation as an attribute-indexed consistency question: agreement on observed conditional feature laws extrapolates to an unseen attribute exactly when the transported attribute-potential identities extend to that attribute.

A.5 Unifying frameworks

Several works abstract away from individual identifiability theorems to identify common proof principles. Khemakhem et al. [2020a] unify nonlinear ICA and identifiable VAEs through condition-dependent exponential-family latent priors, but this unification is tied to conditional factorization and sufficient variability. Ahuja et al. [2022a] give a mechanism-based and equivariance view of identifiable representation learning, characterizing residual ambiguity through symmetries shared by the latent mechanisms. Yao et al. [2025] propose an invariance-principle view of CRL, showing how several CRL guarantees can be understood through representation alignment with environment-induced invariances. Reizinger et al. [2025] introduce identifiable exchangeable mechanisms as a probabilistic framework connecting ICA, CRL, and causal structure learning through exchangeable non-i.i.d. data. Squires and Ravikumar [2026] develop a unified framework for unsupervised concept extraction, where identifiability reduces to showing that the intersection of valid transition sets lies inside the allowable ambiguity class. These frameworks clarify identifiability within particular structural regimes. CMMs are complementary: they provide a conditional framework in which observed feature agreement induces constrained latent concept transitions, and the same attribute-potential constraints characterize extrapolation to unseen attributes.

B Preliminaries

B.1 Notation summary

We provide in Table 1 a summary of notation used throughout the appendix.

Symbol	Clarification
Spaces, attributes, and variables	
$\mathcal{A}, \mathcal{A}_{\mathrm{o}}, \mathcal{A}_{\mathrm{ex}}$	Full attribute space, observed attributes, and target/extrapolation attributes.
A	Attribute random variable.
$(a, a_0, a_{\mathrm{ex}})$	Generic attribute value, anchor attribute in $\mathcal{A}_{\mathrm{o}}$, and generic extrapolation attribute.
$(\mathscr{L}, \Lambda, \lambda)$	Modulator space, modulator random variable, and modulator value.
$(\mathcal{C}, C, c)$	Concept space, concept random variable, and concept value.
$(\mathcal{X}, X, x)$	Feature space, feature random variable, and feature value.
CMMs and induced kernels	
M, M'	Two CMMs, typically $\mathsf{M} = (\mathbf{Q}, \mathbf{B}, \mathbf{K})$ and $\mathsf{M}' = (\mathbf{Q}', \mathbf{B}, \mathbf{K}')$.
$\mathcal{M} = \mathcal{Q} \times \{\mathbf{B}\} \times \mathcal{K}$	CMM class with fixed concept-modulation kernel $\mathbf{B}$.
$(\mathcal{Q}, \mathcal{K})$	Indexing-kernel class and mixing-kernel class.
$(\mathbf{Q}, \mathbf{B}, \mathbf{K})$	Indexing kernel, fixed concept-modulation kernel, and mixing kernel.
$\mathbf{P}^{\mathsf{M}}$	Induced feature-generation kernel $\mathbf{KBQ}$.
$(\bar{\mathbf{Q}}, \bar{\mathcal{Q}})$	Induced concept kernel $\mathbf{BQ}$ and induced concept-kernel class $\mathbf{B}\mathcal{Q}$.
$\mathbf{T}_f$	Pushforward operator induced by a measurable map f.
Reference measures, densities, and potentials	
ν	Reference measure on $\mathcal{C}$.
$p_{\bar{\mathbf{Q}}}(c \mid a)$	Density of $\bar{\mathbf{Q}}(\cdot \mid a)$ with respect to ν.
$\Delta^{\bar{\mathbf{Q}}}_{a,a'}$	Attribute potential, $\Delta^{\bar{\mathbf{Q}}}_{a,a'} := \log p_{\bar{\mathbf{Q}}}(c \mid a) - \log p_{\bar{\mathbf{Q}}}(c \mid a')$.
$\widetilde{\Delta}^{\bar{\mathbf{Q}}}_{a,a_0}(c, c_0)$	Centered attribute potential $\Delta^{\bar{\mathbf{Q}}}_{a,a_0}(c) - \Delta^{\bar{\mathbf{Q}}}_{a,a_0}(c_0)$.
$\overset{\nu}{=}$	Equality of kernels on all ν-dominated input measures.
Transitions and equivalence relations	
$\mathrm{Aut}_\nu(\mathcal{C})$	Measurable concept-space automorphisms preserving ν-null sets in both directions.
τ	Latent concept transition in $\mathrm{Aut}_\nu(\mathcal{C})$.
$J\tau$	Jacobian of the latent concept transition τ.
$r_{\tau^{-1}}$	Radon–Nikodym derivative $d(\mathbf{T}_{\tau^{-1}}\nu)/d\nu$.
$(\mathcal{T}^{\nu}_{\mathbf{K}}, \mathcal{T}^{\nu}_{\bar{\mathbf{Q}}})$	Valid mixing and concept transition sets.
$(\mathfrak{G}, \sim_{\mathfrak{G}})$	Allowed transition group and induced model-equivalence relation.
Extrapolation notation	
φ	Fixed attribute representation used in affine attribute-potential extrapolation.
$D^{\bar{\mathbf{Q}}}$	Coefficient map for centered attribute-potential differences in Thm. 4.

Table 1: Notation summary.

B.2 Measure-theoretic background

We use the measure-theoretic notation introduced in the main text. The only additional convention needed in the appendix concerns restrictions of Markov kernels. For a measurable space $\mathcal{U}$, write $\mathcal{B}(\mathcal{U})$ for its σ-algebra. If $\mathcal{U}' \in \mathcal{B}(\mathcal{U})$, then $\mathcal{U}'$ is equipped with the trace σ-algebra

$$\mathcal{B}(\mathcal{U}') := \{E \cap \mathcal{U}' \mid E \in \mathcal{B}(\mathcal{U})\}.$$

If $\mathbf{K} \in \mathcal{K}^{\mathrm{mk}}(\mathcal{U} \to \mathcal{V})$, then $\mathbf{K}_{\mathcal{U}'} \in \mathcal{K}^{\mathrm{mk}}(\mathcal{U}' \to \mathcal{V})$ denotes the domain restriction of $\mathbf{K}$, defined by

$$\mathbf{K}_{\mathcal{U}'}(B \mid u) := \mathbf{K}(B \mid u), \qquad u \in \mathcal{U}', \qquad B \in \mathcal{B}(\mathcal{V}).$$

This is well defined because, for each $B \in \mathcal{B}(\mathcal{V})$, the map $u \mapsto \mathbf{K}(B \mid u)$ is $\mathcal{B}(\mathcal{U})$-measurable, and its restriction to $\mathcal{U}'$ is $\mathcal{B}(\mathcal{U}')$-measurable. Thus $\mathbf{K}_{\mathcal{U}'}$ restricts only the source measurable space; it does not restrict the target space and involves no renormalization.

C Proofs

C.1 Proof of $\sim_{\mathfrak{G}}$ being equivalence relation

Proposition 1 ($\sim_{\mathfrak{G}}$ is an equivalence relation)**.** *Fix a CMM class $\mathcal{M} = \mathcal{Q} \times \{\mathbf{B}\} \times \mathcal{K}$, an observed attribute set $\mathcal{A}_{\mathrm{o}} \subseteq \mathcal{A}$, and a transition group $\mathfrak{G} \subseteq \mathrm{Aut}_{\nu}(\mathcal{C})$. For $\mathsf{M} = (\mathbf{Q}, \mathbf{B}, \mathbf{K})$ and $\mathsf{M}' = (\mathbf{Q}', \mathbf{B}, \mathbf{K}')$, define $\mathsf{M} \sim_{\mathfrak{G}} \mathsf{M}'$ on $\mathcal{A}_{\mathrm{o}}$ if there exists $\tau \in \mathfrak{G}$ such that*

$$\mathbf{K}' \overset{\nu}{=} \mathbf{K}\mathbf{T}_{\tau}, \qquad \bar{\mathbf{Q}}(\cdot \mid a) = \mathbf{T}_{\tau}\bar{\mathbf{Q}}'(\cdot \mid a) \quad \forall a \in \mathcal{A}_{\mathrm{o}},$$

where $\bar{\mathbf{Q}} = \mathbf{B}\mathbf{Q}$ and $\bar{\mathbf{Q}}' = \mathbf{B}\mathbf{Q}'$. Then $\sim_{\mathfrak{G}}$ is an equivalence relation on $\mathcal{M}$ restricted to $\mathcal{A}_{\mathrm{o}}$.

Proof. Fix a CMM class $\mathcal{M} = \mathcal{Q} \times \{\mathbf{B}\} \times \mathcal{K}$, an observed attribute set $\mathcal{A}_{\mathrm{o}} \subseteq \mathcal{A}$, and a transition group $\mathfrak{G} \subseteq \mathrm{Aut}_{\nu}(\mathcal{C})$. Recall that $\mathsf{M} = (\mathbf{Q}, \mathbf{B}, \mathbf{K})$ and $\mathsf{M}' = (\mathbf{Q}', \mathbf{B}, \mathbf{K}')$ satisfy $\mathsf{M} \sim_{\mathfrak{G}} \mathsf{M}'$ on $\mathcal{A}_{\mathrm{o}}$ if there exists $\tau \in \mathfrak{G}$ such that

$$\mathbf{K}' \overset{\nu}{=} \mathbf{K}\mathbf{T}_{\tau}, \qquad \bar{\mathbf{Q}}(\cdot \mid a) = \mathbf{T}_{\tau}\bar{\mathbf{Q}}'(\cdot \mid a) \quad \forall a \in \mathcal{A}_{\mathrm{o}},$$

where $\bar{\mathbf{Q}} = \mathbf{B}\mathbf{Q}$ and $\bar{\mathbf{Q}}' = \mathbf{B}\mathbf{Q}'$.

We first prove reflexivity. Since $\mathfrak{G}$ is a group, $\mathrm{id}_{\mathcal{C}} \in \mathfrak{G}$. For every $\mathsf{M} = (\mathbf{Q}, \mathbf{B}, \mathbf{K})$, we have $\mathbf{K} \overset{\nu}{=} \mathbf{K}\mathbf{T}_{\mathrm{id}_{\mathcal{C}}}$ and $\bar{\mathbf{Q}}(\cdot \mid a) = \mathbf{T}_{\mathrm{id}_{\mathcal{C}}}\bar{\mathbf{Q}}(\cdot \mid a)$ for all $a \in \mathcal{A}_{\mathrm{o}}$. Hence $\mathsf{M} \sim_{\mathfrak{G}} \mathsf{M}$.

We next prove symmetry. Suppose $\mathsf{M} \sim_{\mathfrak{G}} \mathsf{M}'$ through $\tau \in \mathfrak{G}$. Then $\tau^{-1} \in \mathfrak{G}$. For any $\mu \ll \nu$, since $\tau^{-1} \in \mathrm{Aut}_{\nu}(\mathcal{C})$, we have $\mathbf{T}_{\tau^{-1}}\mu \ll \nu$, and therefore

$$\mathbf{K}'\mathbf{T}_{\tau^{-1}}\mu = \mathbf{K}\mathbf{T}_{\tau}\mathbf{T}_{\tau^{-1}}\mu = \mathbf{K}\mu.$$

Thus $\mathbf{K} \overset{\nu}{=} \mathbf{K}'\mathbf{T}_{\tau^{-1}}$. Similarly, for every $a \in \mathcal{A}_{\mathrm{o}}$,

$$\mathbf{T}_{\tau^{-1}}\bar{\mathbf{Q}}(\cdot \mid a) = \mathbf{T}_{\tau^{-1}}\mathbf{T}_{\tau}\bar{\mathbf{Q}}'(\cdot \mid a) = \bar{\mathbf{Q}}'(\cdot \mid a),$$

with equalities understood modulo ν-null sets. Hence $\mathsf{M}' \sim_{\mathfrak{G}} \mathsf{M}$.

Finally, we prove transitivity. Suppose $\mathsf{M} \sim_{\mathfrak{G}} \mathsf{M}'$ through $\tau \in \mathfrak{G}$, and $\mathsf{M}' \sim_{\mathfrak{G}} \mathsf{M}''$ through $\sigma \in \mathfrak{G}$. Write $\mathsf{M}'' = (\mathbf{Q}'', \mathbf{B}, \mathbf{K}'')$ and $\bar{\mathbf{Q}}'' = \mathbf{B}\mathbf{Q}''$. Since $\mathfrak{G}$ is a group, $\tau \circ \sigma \in \mathfrak{G}$. For every $\mu \ll \nu$, $\mathbf{T}_{\sigma}\mu \ll \nu$, so

$$\mathbf{K}''\mu = \mathbf{K}'\mathbf{T}_{\sigma}\mu = \mathbf{K}\mathbf{T}_{\tau}\mathbf{T}_{\sigma}\mu = \mathbf{K}\mathbf{T}_{\tau\circ\sigma}\mu.$$

Hence $\mathbf{K}'' \overset{\nu}{=} \mathbf{K}\mathbf{T}_{\tau\circ\sigma}$. Moreover, for every $a \in \mathcal{A}_{\mathrm{o}}$,

$$\bar{\mathbf{Q}}(\cdot \mid a) = \mathbf{T}_{\tau}\bar{\mathbf{Q}}'(\cdot \mid a) = \mathbf{T}_{\tau}\mathbf{T}_{\sigma}\bar{\mathbf{Q}}''(\cdot \mid a) = \mathbf{T}_{\tau\circ\sigma}\bar{\mathbf{Q}}''(\cdot \mid a).$$

Thus $\mathsf{M} \sim_{\mathfrak{G}} \mathsf{M}''$. Therefore $\sim_{\mathfrak{G}}$ is reflexive, symmetric, and transitive, and hence is an equivalence relation on $\mathcal{M}$ restricted to $\mathcal{A}_{\mathrm{o}}$. □

C.2 Proof of Thm. 1

Thm. 1 actually only requires the following condition, which is implied by Definition 6 when $\mathcal{A}_{\mathrm{o}} \neq \varnothing$.

Assumption 1. *The conditional concept distributions induced by the CMM class $\mathcal{M} = \mathcal{Q} \times \{\mathbf{B}\} \times \mathcal{K}$ are dominated by the reference measure ν, i.e.,*

$$\mathbf{B}\mathbf{Q}(\cdot \mid a) \ll \nu \qquad \forall \mathbf{Q} \in \mathcal{Q},\ \forall a \in \mathcal{A}.$$

In addition, there exists a probability measure π on $\mathcal{A}_{\mathrm{o}}$ such that

$$\int_{\mathcal{A}_{\mathrm{o}}} \mathbf{B}\mathbf{Q}(\cdot \mid a)\, \pi(da) \sim \nu \qquad \forall \mathbf{Q} \in \mathcal{Q},$$

so that the observed attributes collectively cover the reference measure ν up to its null sets.

We now prove the following theorem, restated for convenience, under Assumption 1.

Theorem 1 (CMM-lift of the intersection theorem)**.** *Let $\mathcal{M} = \mathcal{Q} \times \{\mathbf{B}\} \times \mathcal{K}$ be a CMM class, and let $\mathcal{A}_{\mathrm{o}} \neq \varnothing$. Assume that $\mathcal{K}$ is ν-Blackwell reducible and that the induced concept-kernel class $\bar{\mathcal{Q}} = \mathbf{B}\mathcal{Q}$ has common ν-support. Then, for any $\mathsf{M} = (\mathbf{Q}, \mathbf{B}, \mathbf{K}), \mathsf{M}' = (\mathbf{Q}', \mathbf{B}, \mathbf{K}') \in \mathcal{M}$ feature-equivalent on $\mathcal{A}_{\mathrm{o}}$, with $\bar{\mathbf{Q}} = \mathbf{B}\mathbf{Q}$, there exists*

$$\tau \in \mathcal{T}^{\nu}_{\mathbf{K}}(\mathcal{K}) \cap \mathcal{T}^{\nu}_{\bar{\mathbf{Q}}}(\bar{\mathcal{Q}}; \mathcal{A}_{\mathrm{o}})$$

such that M and M' are Blackwell-equivalent through τ on $\mathcal{A}_{\mathrm{o}}$. If, moreover, $\mathcal{T}^{\nu}_{\mathbf{K}}(\mathcal{K}) \cap \mathcal{T}^{\nu}_{\bar{\mathbf{Q}}}(\bar{\mathcal{Q}}; \mathcal{A}_{\mathrm{o}}) \subseteq \mathfrak{G}$ for a transition group $\mathfrak{G} \subseteq \mathrm{Aut}_{\nu}(\mathcal{C})$, then M is identifiable up to $\sim_{\mathfrak{G}}$ in $\mathcal{M}$ on $\mathcal{A}_{\mathrm{o}}$, i.e., every $\mathsf{M}' \in \mathcal{M}$ feature-equivalent to M on $\mathcal{A}_{\mathrm{o}}$ satisfies $\mathsf{M} \sim_{\mathfrak{G}} \mathsf{M}'$.

Proof. Let $\mathsf{M}' = (\mathbf{Q}', \mathbf{B}, \mathbf{K}') \in \mathcal{M}$ be feature-equivalent to $\mathsf{M} = (\mathbf{Q}, \mathbf{B}, \mathbf{K})$ on $\mathcal{A}_{\mathrm{o}}$. Write

$$\mu_a := \mathbf{B}\mathbf{Q}(\cdot \mid a), \qquad \mu'_a := \mathbf{B}\mathbf{Q}'(\cdot \mid a),$$

which satisfies $\mu_a, \mu'_a \ll \nu$ for all $a \in \mathcal{A}_{\mathrm{o}}$ by Assumption 1. By ν-Blackwell reducibility, choose bimeasurable embeddings $g, g' \in \mathcal{G}$ such that

$$\mathbf{K} \overset{\nu}{=} \widetilde{\mathbf{K}}\mathbf{T}_g, \qquad \mathbf{K}' \overset{\nu}{=} \widetilde{\mathbf{K}}\mathbf{T}_{g'}.$$

Feature equivalence and carrier equality give, for every $a \in \mathcal{A}_{\mathrm{o}}$,

$$\widetilde{\mathbf{K}}\mathbf{T}_g\mu_a = \widetilde{\mathbf{K}}\mathbf{T}_{g'}\mu'_a.$$

Since $\widetilde{\mathbf{K}}$ is injective on the relevant pushed-forward carrier,

$$\mathbf{T}_g\mu_a = \mathbf{T}_{g'}\mu'_a \qquad \forall a \in \mathcal{A}_{\mathrm{o}}.$$

Let π be the full-carrier anchor from Assumption 1, and define

$$\overline{\mu}_{\mathbf{Q}} := \int_{\mathcal{A}_{\mathrm{o}}} \mu_a \, \pi(da), \qquad \overline{\mu}_{\mathbf{Q}'} := \int_{\mathcal{A}_{\mathrm{o}}} \mu'_a \, \pi(da).$$

Integrating the previous display over $a \sim \pi$ gives

$$\mathbf{T}_g\overline{\mu}_{\mathbf{Q}} = \mathbf{T}_{g'}\overline{\mu}_{\mathbf{Q}'}.$$

Since $\overline{\mu}_{\mathbf{Q}} \sim \nu$ and $\overline{\mu}_{\mathbf{Q}'} \sim \nu$, it follows that

$$\mathbf{T}_g\nu \sim \mathbf{T}_{g'}\nu.$$

Hence $g(\mathcal{C})$ and $g'(\mathcal{C})$ agree modulo the corresponding pushed-forward carrier:

$$\nu\{c : g'(c) \notin g(\mathcal{C})\} = 0, \qquad \nu\{c : g(c) \notin g'(\mathcal{C})\} = 0.$$

Therefore the maps

$$\tau := g^{-1} \circ g', \qquad \tau^{-1} := g'^{-1} \circ g$$

are well-defined modulo ν-null sets. Moreover, since $\mathbf{T}_g\mathbf{T}_\tau\nu = \mathbf{T}_{g'}\nu \sim \mathbf{T}_g\nu$, injectivity of g implies $\mathbf{T}_\tau\nu \sim \nu$. The same argument with g and g' exchanged gives $\mathbf{T}_{\tau^{-1}}\nu \sim \nu$. Thus

$$\tau \in \mathrm{Aut}_{\nu}(\mathcal{C}), \qquad g' = g \circ \tau \quad \nu\text{-a.e.}$$

Since $\mu'_a \ll \nu$, the relation $g' = g \circ \tau$ ν-a.e. implies

$$\mathbf{T}_{g'}\mu'_a = \mathbf{T}_g\mathbf{T}_\tau\mu'_a.$$

Combining this with $\mathbf{T}_g\mu_a = \mathbf{T}_{g'}\mu'_a$, we obtain

$$\mathbf{T}_g\mu_a = \mathbf{T}_g\mathbf{T}_\tau\mu'_a.$$

Since g is a bimeasurable embedding, $\mathbf{T}_g$ is injective on measures on $\mathcal{C}$. Hence

$$\mu_a = \mathbf{T}_\tau\mu'_a \qquad \forall a \in \mathcal{A}_{\mathrm{o}}.$$

Equivalently,

$$\mathbf{B}\mathbf{Q}(\cdot \mid a) = \mathbf{T}_\tau\mathbf{B}\mathbf{Q}'(\cdot \mid a) \qquad \forall a \in \mathcal{A}_{\mathrm{o}},$$

so $\tau \in \mathcal{T}^{\nu}_{\bar{\mathbf{Q}}}(\bar{\mathcal{Q}}; \mathcal{A}_{\mathrm{o}})$.

Finally, for every $\mu \ll \nu$,

$$\mathbf{K}'\mu = \widetilde{\mathbf{K}}\mathbf{T}_{g'}\mu = \widetilde{\mathbf{K}}\mathbf{T}_g\mathbf{T}_\tau\mu = \mathbf{K}\mathbf{T}_\tau\mu.$$

Thus $\mathbf{K}' \overset{\nu}{=} \mathbf{K}\mathbf{T}_\tau$, so $\tau \in \mathcal{T}^{\nu}_{\mathbf{K}}(\mathcal{K})$. Therefore

$$\tau \in \mathcal{T}^{\nu}_{\mathbf{K}}(\mathcal{K}) \cap \mathcal{T}^{\nu}_{\bar{\mathbf{Q}}}(\bar{\mathcal{Q}}; \mathcal{A}_{\mathrm{o}}) \subseteq \mathfrak{G}.$$

By Definition 5, $\mathsf{M} \sim_{\mathfrak{G}} \mathsf{M}'$. Since M' was arbitrary among feature-equivalent alternatives, M is identifiable up to $\sim_{\mathfrak{G}}$. □

C.3 Proof of Thm. 2

Theorem 2 (Density characterization of concept transitions)**.** *Consider a fixed concept modulation kernel* $\mathbf{B}$*, and assume that* $\bar{\mathcal{Q}} = \mathbf{B}\mathcal{Q}$ *has common* ν*-support as defined in Definition 6. Let* $\mathbf{Q}, \mathbf{Q}' \in \mathcal{Q}$ *be two indexing kernels. Let* $\mathcal{A}' \subseteq \mathcal{A}$*,* $\tau \in \mathrm{Aut}_\nu(\mathcal{C})$*, and define* $r_{\tau^{-1}} := d(\mathbf{T}_{\tau^{-1}}\nu)/d\nu$*. For any fixed anchor* $a_0 \in \mathcal{A}'$ *and* $\bar{\mathbf{Q}} = \mathbf{B}\mathbf{Q}$*,* $\bar{\mathbf{Q}}' = \mathbf{B}\mathbf{Q}'$*, the following are equivalent:*

(i) $\bar{\mathbf{Q}}_{\mathcal{A}'} = \mathbf{T}_\tau \bar{\mathbf{Q}}'_{\mathcal{A}'}$.

(ii) For every $a \in \mathcal{A}'$*,* $p_{\bar{\mathbf{Q}}'}(c \mid a) = p_{\bar{\mathbf{Q}}}(\tau(c) \mid a) r_{\tau^{-1}}(c)$ *holds for* ν*-a.e.* c*.*

(iii) The anchored density identity $p_{\bar{\mathbf{Q}}'}(c \mid a_0) = p_{\bar{\mathbf{Q}}}(\tau(c) \mid a_0) r_{\tau^{-1}}(c)$ *holds for* ν*-a.e.* c*, and for every* $a \in \mathcal{A}'$*,* $\Delta^{\bar{\mathbf{Q}}'}_{a,a_0}(c) = \Delta^{\bar{\mathbf{Q}}}_{a,a_0}(\tau(c))$ *holds* ν*-a.e.* c*.*

Consequently, $\tau \in \mathcal{T}^\nu_{\bar{\mathbf{Q}}}(\bar{\mathcal{Q}}; \mathcal{A}_\mathrm{o})$ *if and only if there exists* $\mathbf{Q}' \in \mathcal{Q}$ *such that these equivalent conditions hold with* $\mathcal{A}' = \mathcal{A}_\mathrm{o}$*.*

Proof. We first show that (i) and (ii) are equivalent. For any $a \in \mathcal{A}'$, the identity $\bar{\mathbf{Q}}(\cdot \mid a) = \mathbf{T}_\tau \bar{\mathbf{Q}}'(\cdot \mid a)$ is equivalent, since $\tau \in \mathrm{Aut}_\nu(\mathcal{C})$, to

$$\bar{\mathbf{Q}}'(\cdot \mid a) = \mathbf{T}_{\tau^{-1}} \bar{\mathbf{Q}}(\cdot \mid a).$$

If $\bar{\mathbf{Q}}(\cdot \mid a)$ has density $p_{\bar{\mathbf{Q}}}(\cdot \mid a)$ with respect to ν, then the Radon–Nikodym chain rule for $\mathbf{T}_{\tau^{-1}}$ gives

$$\frac{d\,\mathbf{T}_{\tau^{-1}} \bar{\mathbf{Q}}(\cdot \mid a)}{d\nu}(c) = p_{\bar{\mathbf{Q}}}(\tau(c) \mid a) r_{\tau^{-1}}(c), \qquad r_{\tau^{-1}} = \frac{d(\mathbf{T}_{\tau^{-1}}\nu)}{d\nu}.$$

Thus $\bar{\mathbf{Q}}'(\cdot \mid a) = \mathbf{T}_{\tau^{-1}} \bar{\mathbf{Q}}(\cdot \mid a)$ holds for every $a \in \mathcal{A}'$ if and only if

$$p_{\bar{\mathbf{Q}}'}(c \mid a) = p_{\bar{\mathbf{Q}}}(\tau(c) \mid a) r_{\tau^{-1}}(c)$$

holds for every $a \in \mathcal{A}'$ and ν-a.e. c, which proves the equivalence between (i) and (ii).

We next show that (ii) implies (iii). Taking $a = a_0$ in (ii) gives the anchored density identity

$$p_{\bar{\mathbf{Q}}'}(c \mid a_0) = p_{\bar{\mathbf{Q}}}(\tau(c) \mid a_0) r_{\tau^{-1}}(c)$$

for ν-a.e. c. Now fix any $a \in \mathcal{A}'$. By the common-support assumption, $p_{\bar{\mathbf{Q}}'}(\cdot \mid a)$, $p_{\bar{\mathbf{Q}}'}(\cdot \mid a_0)$, $p_{\bar{\mathbf{Q}}}(\cdot \mid a)$, and $p_{\bar{\mathbf{Q}}}(\cdot \mid a_0)$ are positive and finite ν-a.e.; since $\tau \in \mathrm{Aut}_\nu(\mathcal{C})$, the same is true of $p_{\bar{\mathbf{Q}}}(\tau(\cdot) \mid a)$ and $p_{\bar{\mathbf{Q}}}(\tau(\cdot) \mid a_0)$ on a ν-full set. Intersecting this full-measure set with the full-measure sets on which (ii) holds for a and a_0, we may take logarithms and subtract:

$$\begin{aligned}
\log p_{\bar{\mathbf{Q}}'}(c \mid a) - \log p_{\bar{\mathbf{Q}}'}(c \mid a_0) &= \big[\log p_{\bar{\mathbf{Q}}}(\tau(c) \mid a) + \log r_{\tau^{-1}}(c)\big] \\
&\quad - \big[\log p_{\bar{\mathbf{Q}}}(\tau(c) \mid a_0) + \log r_{\tau^{-1}}(c)\big] \\
&= \log p_{\bar{\mathbf{Q}}}(\tau(c) \mid a) - \log p_{\bar{\mathbf{Q}}}(\tau(c) \mid a_0).
\end{aligned}$$

Therefore

$$\Delta^{\bar{\mathbf{Q}}'}_{a,a_0}(c) = \Delta^{\bar{\mathbf{Q}}}_{a,a_0}(\tau(c))$$

for ν-a.e. c. Since $a \in \mathcal{A}'$ was arbitrary, the attribute-potential identities hold for every $a \in \mathcal{A}'$, so (iii) follows.

Conversely, suppose (iii) holds. Then for every $a \in \mathcal{A}'$,

$$\begin{aligned}
p_{\bar{\mathbf{Q}}'}(c \mid a) &= p_{\bar{\mathbf{Q}}'}(c \mid a_0) \exp\{\Delta^{\bar{\mathbf{Q}}'}_{a,a_0}(c)\} \\
&= p_{\bar{\mathbf{Q}}}(\tau(c) \mid a_0) r_{\tau^{-1}}(c) \exp\{\Delta^{\bar{\mathbf{Q}}}_{a,a_0}(\tau(c))\} \\
&= p_{\bar{\mathbf{Q}}}(\tau(c) \mid a) r_{\tau^{-1}}(c)
\end{aligned}$$

for ν-a.e. c, so (ii) holds. The final statement follows from the definition of $\mathcal{T}^\nu_{\bar{\mathbf{Q}}}(\bar{\mathcal{Q}}; \mathcal{A}_\mathrm{o})$. □

C.4 Proof of Thm. 3

Theorem 3 (Attribute-potential characterization of extrapolation)**.** *Let $\mathcal{M} = \mathcal{Q} \times \{\mathbf{B}\} \times \mathcal{K}$ be a CMM class, and let $\mathcal{A}_o \subseteq \mathcal{A}_{ex} \subseteq \mathcal{A}$. Assume that $\mathcal{K}$ is ν-Blackwell reducible and that $\bar{\mathcal{Q}} = \mathbf{B}\mathcal{Q}$ has common ν-support. Fix $a_0 \in \mathcal{A}_o$, and let $\mathsf{M} = (\mathbf{Q}, \mathbf{B}, \mathbf{K})$ and $\mathsf{M}' = (\mathbf{Q}', \mathbf{B}, \mathbf{K}')$ be feature-equivalent on $\mathcal{A}_o$, and let τ be a transition induced by Thm. 1, so that $\mathbf{K}' \overset{\nu}{=} \mathbf{K}\mathbf{T}_\tau$ and $\bar{\mathbf{Q}}_{\mathcal{A}_o} = \mathbf{T}_\tau \bar{\mathbf{Q}}'_{\mathcal{A}_o}$ where $\bar{\mathbf{Q}} = \mathbf{B}\mathbf{Q}$, $\bar{\mathbf{Q}}' = \mathbf{B}\mathbf{Q}'$. Then, for this transition τ,*

$$\mathbf{P}^{\mathsf{M}}_{\mathcal{A}_{ex}} = \mathbf{P}^{\mathsf{M}'}_{\mathcal{A}_{ex}} \quad \Longleftrightarrow \quad \Delta^{\bar{\mathbf{Q}}'}_{a,a_0}(c) = \Delta^{\bar{\mathbf{Q}}}_{a,a_0}(\tau(c)) \text{ for every } a \in \mathcal{A}_{ex},\ \nu\text{-a.e. } c.$$

Thus, if the identities $\Delta^{\bar{\mathbf{Q}}'}_{a,a_0}(c) = \Delta^{\bar{\mathbf{Q}}}_{a,a_0}(\tau(c))$ that hold on $\mathcal{A}_o$ extend over $\mathcal{A}_{ex}$, then $\mathbf{P}^{\mathsf{M}}_{\mathcal{A}_{ex}} = \mathbf{P}^{\mathsf{M}'}_{\mathcal{A}_{ex}}$.

Proof. Write

$$\bar{\mathbf{Q}} = \mathbf{B}\mathbf{Q}, \qquad \bar{\mathbf{Q}}' = \mathbf{B}\mathbf{Q}', \qquad \mu_a := \bar{\mathbf{Q}}(\cdot \mid a), \qquad \mu'_a := \bar{\mathbf{Q}}'(\cdot \mid a).$$

By the common ν-support assumption, $\mu_a, \mu'_a \ll \nu$ for every $a \in \mathcal{A}$. Since $\tau \in \mathrm{Aut}_\nu(\mathcal{C})$, we also have $\mathbf{T}_\tau \mu'_a \ll \nu$ for every $a \in \mathcal{A}$.

The transition τ induced by Thm. 1 satisfies

$$\mathbf{K}' \overset{\nu}{=} \mathbf{K}\mathbf{T}_\tau, \qquad \mu_a = \mathbf{T}_\tau \mu'_a \quad \forall a \in \mathcal{A}_o.$$

In particular,

$$\mu_{a_0} = \mathbf{T}_\tau \mu'_{a_0}. \tag{3}$$

We first show that, for each $a \in \mathcal{A}_{ex}$,

$$\mathbf{P}^{\mathsf{M}}(\cdot \mid a) = \mathbf{P}^{\mathsf{M}'}(\cdot \mid a) \quad \Longleftrightarrow \quad \mu_a = \mathbf{T}_\tau \mu'_a. \tag{4}$$

Indeed, since $\mu'_a \ll \nu$ and $\mathbf{K}' \overset{\nu}{=} \mathbf{K}\mathbf{T}_\tau$,

$$\mathbf{P}^{\mathsf{M}'}(\cdot \mid a) = \mathbf{K}'\mu'_a = \mathbf{K}\mathbf{T}_\tau \mu'_a.$$

Also $\mathbf{P}^{\mathsf{M}}(\cdot \mid a) = \mathbf{K}\mu_a$. Thus feature equality at a is equivalent to

$$\mathbf{K}\mu_a = \mathbf{K}\mathbf{T}_\tau \mu'_a.$$

It remains only to justify that $\mathbf{K}$ is injective on measures dominated by ν. By ν-Blackwell reducibility, choose $g \in \mathcal{G}$ and $\widetilde{\mathbf{K}}$ such that

$$\mathbf{K} \overset{\nu}{=} \widetilde{\mathbf{K}}\mathbf{T}_g.$$

If $\alpha, \beta \ll \nu$ and $\mathbf{K}\alpha = \mathbf{K}\beta$, then

$$\widetilde{\mathbf{K}}\mathbf{T}_g\alpha = \widetilde{\mathbf{K}}\mathbf{T}_g\beta.$$

By the injectivity condition in Definition 3,

$$\mathbf{T}_g\alpha = \mathbf{T}_g\beta.$$

Since g is a bimeasurable embedding, $\mathbf{T}_g$ is injective on probability measures on $\mathcal{C}$, hence $\alpha = \beta$. Applying this with $\alpha = \mu_a$ and $\beta = \mathbf{T}_\tau \mu'_a$ proves (4).

Taking (4) for all $a \in \mathcal{A}_{ex}$ gives

$$\mathbf{P}^{\mathsf{M}}_{\mathcal{A}_{ex}} = \mathbf{P}^{\mathsf{M}'}_{\mathcal{A}_{ex}} \quad \Longleftrightarrow \quad \bar{\mathbf{Q}}_{\mathcal{A}_{ex}} = \mathbf{T}_\tau \bar{\mathbf{Q}}'_{\mathcal{A}_{ex}}. \tag{5}$$

We now translate the right-hand side of (5) into attribute potentials. Apply Thm. 2 with $\mathcal{A}' = \mathcal{A}_{ex}$. It states that

$$\bar{\mathbf{Q}}_{\mathcal{A}_{ex}} = \mathbf{T}_\tau \bar{\mathbf{Q}}'_{\mathcal{A}_{ex}}$$

is equivalent to the conjunction of the anchored density identity

$$p_{\bar{\mathbf{Q}}'}(c \mid a_0) = p_{\bar{\mathbf{Q}}}(\tau(c) \mid a_0) r_{\tau^{-1}}(c) \quad \nu\text{-a.e. } c \tag{6}$$

and the transported attribute-potential identities

$$\Delta^{\bar{\mathbf{Q}}'}_{a,a_0}(c) = \Delta^{\bar{\mathbf{Q}}}_{a,a_0}(\tau(c)) \quad \text{for every } a \in \mathcal{A}_{ex},\ \nu\text{-a.e. } c. \tag{7}$$

However, (6) already follows from (3) by the density form of Thm. 2, since $a_0 \in \mathcal{A}_{\text{o}}$. Therefore, for the fixed transition τ induced from the observed attributes, the condition

$$\bar{\mathbf{Q}}_{\mathcal{A}_{\text{ex}}} = \mathbf{T}_\tau \bar{\mathbf{Q}}'_{\mathcal{A}_{\text{ex}}}$$

is equivalent to (7) alone.

Combining this equivalence with (5), we obtain

$$\mathbf{P}^{\mathsf{M}}_{\mathcal{A}_{\text{ex}}} = \mathbf{P}^{\mathsf{M}'}_{\mathcal{A}_{\text{ex}}} \quad \Longleftrightarrow \quad \Delta^{\bar{\mathbf{Q}}'}_{a,a_0}(c) = \Delta^{\bar{\mathbf{Q}}}_{a,a_0}(\tau(c)) \text{ for every } a \in \mathcal{A}_{\text{ex}},\ \nu\text{-a.e. } c.$$

This proves the claimed equivalence.

The final statement follows immediately: if the transported attribute-potential identities that hold on $\mathcal{A}_{\text{o}}$ extend to all $a \in \mathcal{A}_{\text{ex}}$, then the right-hand side above holds, and therefore $\mathbf{P}^{\mathsf{M}}_{\mathcal{A}_{\text{ex}}} = \mathbf{P}^{\mathsf{M}'}_{\mathcal{A}_{\text{ex}}}$. □

C.5 Proof of Thm. 4

Theorem 4 (Attribute-potential preservation under affine attribute-potential differences)**.** *Let $\mathcal{M} = \mathcal{Q} \times \{\mathbf{B}\} \times \mathcal{K}$ be a CMM class satisfying the assumptions of Thms. 1 and 2. Let $\mathcal{A}_{\text{o}} \subseteq \mathcal{A}_{\text{ex}} \subseteq \mathcal{A}$, and fix $a_0 \in \mathcal{A}_{\text{o}}$. Suppose there exists a fixed attribute representation $\varphi\colon \mathcal{A} \to \mathbb{R}^m$ such that, for every $\bar{\mathbf{Q}} \in \bar{\mathcal{Q}} = \mathbf{B}\mathcal{Q}$, there exists a measurable map $D^{\bar{\mathbf{Q}}}\colon \mathcal{C} \times \mathcal{C} \to \mathbb{R}^m$ satisfying*

$$\Delta^{\bar{\mathbf{Q}}}_{a,a_0}(c) - \Delta^{\bar{\mathbf{Q}}}_{a,a_0}(c_0) = \langle \varphi(a) - \varphi(a_0), D^{\bar{\mathbf{Q}}}(c, c_0) \rangle$$

for every $a \in \mathcal{A}_{\text{ex}}$ and $\nu \otimes \nu$-a.e. (c, c_0). Let $\mathsf{M} = (\mathbf{Q}, \mathbf{B}, \mathbf{K})$ and $\mathsf{M}' = (\mathbf{Q}', \mathbf{B}, \mathbf{K}')$ be feature-equivalent on $\mathcal{A}_{\text{o}}$, and let τ be the transition induced by Thm. 1. If $\varphi(\mathcal{A}_{\text{ex}})$ is contained in the affine hull of $\varphi(\mathcal{A}_{\text{o}})$, then the transported attribute-potential identities extend from $\mathcal{A}_{\text{o}}$ to $\mathcal{A}_{\text{ex}}$: for every $a \in \mathcal{A}_{\text{ex}}$, $\Delta^{\bar{\mathbf{Q}}'}_{a,a_0}(c) = \Delta^{\bar{\mathbf{Q}}}_{a,a_0}(\tau(c))$ for ν-a.e. c.

Proof. Fix $a_{\text{ex}} \in \mathcal{A}_{\text{ex}}$. Since $\varphi(\mathcal{A}_{\text{ex}})$ is contained in the affine hull of $\varphi(\mathcal{A}_{\text{o}})$, there exist $a_1, \dots, a_\ell \in \mathcal{A}_{\text{o}}$ and $\alpha_1, \dots, \alpha_\ell \in \mathbb{R}$ such that $\varphi(a_{\text{ex}}) = \sum_{i=1}^{\ell} \alpha_i \varphi(a_i)$ and $\sum_{i=1}^{\ell} \alpha_i = 1$. By the assumed affine representation of attribute-potential differences, for $\mathsf{M} = (\mathbf{Q}, \mathbf{B}, \mathbf{K})$ we have

$$\begin{aligned}
\Delta^{\bar{\mathbf{Q}}}_{a_{\text{ex}},a_0}(c) - \Delta^{\bar{\mathbf{Q}}}_{a_{\text{ex}},a_0}(c_0) &= \langle \varphi(a_{\text{ex}}) - \varphi(a_0), D^{\bar{\mathbf{Q}}}(c, c_0) \rangle \\
&= \langle \sum_{i=1}^{\ell} \alpha_i \varphi(a_i) - \sum_{i=1}^{\ell} \alpha_i \varphi(a_0), D^{\bar{\mathbf{Q}}}(c, c_0) \rangle \\
&= \sum_{i=1}^{\ell} \alpha_i \langle \varphi(a_i) - \varphi(a_0), D^{\bar{\mathbf{Q}}}(c, c_0) \rangle \\
&= \sum_{i=1}^{\ell} \alpha_i \left[\Delta^{\bar{\mathbf{Q}}}_{a_i,a_0}(c) - \Delta^{\bar{\mathbf{Q}}}_{a_i,a_0}(c_0) \right],
\end{aligned}$$

where the second equality uses $\sum_{i=1}^{\ell} \alpha_i = 1$. The same argument for $\mathsf{M}' = (\mathbf{Q}', \mathbf{B}, \mathbf{K}')$ gives

$$\Delta^{\bar{\mathbf{Q}}'}_{a_{\text{ex}},a_0}(c) - \Delta^{\bar{\mathbf{Q}}'}_{a_{\text{ex}},a_0}(c_0) = \sum_{i=1}^{\ell} \alpha_i \left[\Delta^{\bar{\mathbf{Q}}'}_{a_i,a_0}(c) - \Delta^{\bar{\mathbf{Q}}'}_{a_i,a_0}(c_0) \right].$$

Since $a_i \in \mathcal{A}_{\text{o}}$, Thm. 2 gives the observed transported identities

$$\Delta^{\bar{\mathbf{Q}}'}_{a_i,a_0}(c) = \Delta^{\bar{\mathbf{Q}}}_{a_i,a_0}(\tau(c)) \qquad \text{for } \nu\text{-a.e. } c$$

for each $i \in [\ell]$. Subtracting the identity evaluated at c_0 from the identity evaluated at c, we obtain

$$\Delta^{\bar{\mathbf{Q}}'}_{a_i,a_0}(c) - \Delta^{\bar{\mathbf{Q}}'}_{a_i,a_0}(c_0) = \Delta^{\bar{\mathbf{Q}}}_{a_i,a_0}(\tau(c)) - \Delta^{\bar{\mathbf{Q}}}_{a_i,a_0}(\tau(c_0))$$

for $\nu \otimes \nu$-a.e. (c, c_0), given that $\mathbf{T}_\tau \nu \sim \nu$. Combining the preceding two displays gives

$$\Delta^{\bar{\mathbf{Q}}'}_{a_{\text{ex}},a_0}(c) - \Delta^{\bar{\mathbf{Q}}'}_{a_{\text{ex}},a_0}(c_0) = \sum_{i=1}^{\ell} \alpha_i \left[\Delta^{\bar{\mathbf{Q}}}_{a_i,a_0}(\tau(c)) - \Delta^{\bar{\mathbf{Q}}}_{a_i,a_0}(\tau(c_0)) \right].$$

On the other hand, applying the affine representation for M at the pair $(\tau(c), \tau(c_0))$ gives

$$\Delta^{\bar{\mathbf{Q}}}_{a_{\text{ex}},a_0}(\tau(c)) - \Delta^{\bar{\mathbf{Q}}}_{a_{\text{ex}},a_0}(\tau(c_0)) = \sum_{i=1}^{\ell} \alpha_i \left[\Delta^{\bar{\mathbf{Q}}}_{a_i,a_0}(\tau(c)) - \Delta^{\bar{\mathbf{Q}}}_{a_i,a_0}(\tau(c_0))\right]$$

for $\nu \otimes \nu$-a.e. (c, c_0), since $\tau \in \mathrm{Aut}_\nu(\mathcal{C})$ preserves ν-null sets. Therefore

$$\Delta^{\bar{\mathbf{Q}}'}_{a_{\text{ex}},a_0}(c) - \Delta^{\bar{\mathbf{Q}}'}_{a_{\text{ex}},a_0}(c_0) = \Delta^{\bar{\mathbf{Q}}}_{a_{\text{ex}},a_0}(\tau(c)) - \Delta^{\bar{\mathbf{Q}}}_{a_{\text{ex}},a_0}(\tau(c_0))$$

for $\nu \otimes \nu$-a.e. (c, c_0). Define

$$H(c) := \Delta^{\bar{\mathbf{Q}}'}_{a_{\text{ex}},a_0}(c) - \Delta^{\bar{\mathbf{Q}}}_{a_{\text{ex}},a_0}(\tau(c)).$$

The preceding display says $H(c) - H(c_0) = 0$ for $\nu \otimes \nu$-a.e. (c, c_0). This implies that, through Fubini, H is ν-a.e. constant, so there exists a scalar $b(a_{\text{ex}})$ such that

$$\Delta^{\bar{\mathbf{Q}}'}_{a_{\text{ex}},a_0}(c) = \Delta^{\bar{\mathbf{Q}}}_{a_{\text{ex}},a_0}(\tau(c)) + b(a_{\text{ex}})$$

for ν-a.e. c. Let $r_{\tau^{-1}}(c) := d(\mathbf{T}_{\tau^{-1}}\nu)/d\nu(c)$. By the anchored density identity in Thm. 2,

$$p_{\bar{\mathbf{Q}}'}(c \mid a_0) = p_{\bar{\mathbf{Q}}}(\tau(c) \mid a_0) r_{\tau^{-1}}(c)$$

for ν-a.e. c. Therefore,

$$\begin{aligned} p_{\bar{\mathbf{Q}}'}(c \mid a_{\text{ex}}) &= p_{\bar{\mathbf{Q}}'}(c \mid a_0) \exp\{\Delta^{\bar{\mathbf{Q}}'}_{a_{\text{ex}},a_0}(c)\} \\ &= p_{\bar{\mathbf{Q}}}(\tau(c) \mid a_0) r_{\tau^{-1}}(c) \exp\{\Delta^{\bar{\mathbf{Q}}}_{a_{\text{ex}},a_0}(\tau(c)) + b(a_{\text{ex}})\} \\ &= \exp\{b(a_{\text{ex}})\} p_{\bar{\mathbf{Q}}}(\tau(c) \mid a_{\text{ex}}) r_{\tau^{-1}}(c). \end{aligned}$$

Integrating both sides with respect to ν gives

$$\begin{aligned} 1 &= \exp\{b(a_{\text{ex}})\} \int p_{\bar{\mathbf{Q}}}(\tau(c) \mid a_{\text{ex}}) r_{\tau^{-1}}(c)\, \nu(dc) \\ &= \exp\{b(a_{\text{ex}})\} \int p_{\bar{\mathbf{Q}}}(u \mid a_{\text{ex}})\, \nu(du) \\ &= \exp\{b(a_{\text{ex}})\}. \end{aligned}$$

Hence $b(a_{\text{ex}}) = 0$. Thus $\Delta^{\bar{\mathbf{Q}}'}_{a_{\text{ex}},a_0}(c) = \Delta^{\bar{\mathbf{Q}}}_{a_{\text{ex}},a_0}(\tau(c))$ for ν-a.e. c. Since $a_{\text{ex}} \in \mathcal{A}_{\text{ex}}$ was arbitrary, the transported attribute-potential identities extend from $\mathcal{A}_{\text{o}}$ to $\mathcal{A}_{\text{ex}}$. □

C.6 Proof of Cor. 1

Corollary 1 (Interaction-basis extrapolation)**.** *Let $\mathcal{M} = \mathcal{Q} \times \{\mathbf{B}\} \times \mathcal{K}$ satisfy the assumptions of Thm. 3. Let $\mathcal{A} = 2^{[m]}$, let $\mathcal{H} \subseteq 2^{[m]}$ be a finite downward-closed family containing $\varnothing$, i.e., $H \in \mathcal{H}$ and $U \subseteq H$ imply $U \in \mathcal{H}$, and set $\mathcal{A}_{\text{o}} = \mathcal{H}$. Assume every $\mathsf{M} \in \mathcal{M}$ induces concept densities of the form $p^{\mathsf{M}}(c \mid S) = \exp\left(\sum_{T \in \mathcal{H},\, T \subseteq S} h^{\mathsf{M}}_T(c)\right) / Z^{\mathsf{M}}(S)$. Then feature equivalence on $\mathcal{A}_{\text{o}}$ implies feature equivalence on all $S \subseteq [m]$.*

Proof. Set $\mathcal{A}_{\text{ex}} = 2^{[m]}$ and take $a_0 = \varnothing$. For each $S \subseteq [m]$, define the interaction-incidence vector $\varphi_{\mathcal{H}}(S) \in \mathbb{R}^{\mathcal{H}}$ by $\varphi_{\mathcal{H}}(S)_T = \mathbf{1}\{T \subseteq S\}$ for $T \in \mathcal{H}$. We first show that the assumed density model has the affine attribute-potential form required by Thm. 4. Write

$$\ell^{\mathsf{M}}_S(c) = \sum_{T \in \mathcal{H}, T \subseteq S} h^{\mathsf{M}}_T(c),$$

so that $p^{\mathsf{M}}(c \mid S) = \exp\{\ell^{\mathsf{M}}_S(c)\}/Z^{\mathsf{M}}_S$ for a normalizing constant Z^{M}_S. Since $p^{\mathsf{M}}(c \mid \varnothing) = \exp\{h^{\mathsf{M}}_\varnothing(c)\}/Z^{\mathsf{M}}_\varnothing$, the attribute potential relative to $\varnothing$ is

$$\Delta^{\mathsf{M}}_{S,\varnothing}(c) = \log \frac{p^{\mathsf{M}}(c \mid S)}{p^{\mathsf{M}}(c \mid \varnothing)} = \sum_{T \in \mathcal{H}, T \subseteq S, T \neq \varnothing} h^{\mathsf{M}}_T(c) - \log \frac{Z^{\mathsf{M}}_S}{Z^{\mathsf{M}}_\varnothing}.$$

The normalizing constant does not depend on c, so it disappears after centering at any fixed $c_0 \in \mathcal{C}$. Thus

$$\Delta^{\mathsf{M}}_{S,\varnothing}(c) - \Delta^{\mathsf{M}}_{S,\varnothing}(c_0) = \sum_{T\in\mathcal{H},T\subseteq S,T\neq\varnothing} \big(h^{\mathsf{M}}_T(c) - h^{\mathsf{M}}_T(c_0)\big).$$

Define $D^{\mathsf{M}}(c, c_0) \in \mathbb{R}^{\mathcal{H}}$ by $D^{\mathsf{M}}_{\varnothing}(c, c_0) = 0$ and $D^{\mathsf{M}}_T(c, c_0) = h^{\mathsf{M}}_T(c) - h^{\mathsf{M}}_T(c_0)$ for $T \neq \varnothing$. Since $\varphi_{\mathcal{H}}(\varnothing)$ is equal to 1 in the $\varnothing$-coordinate and 0 elsewhere, the previous display can be written as

$$\Delta^{\mathsf{M}}_{S,\varnothing}(c) - \Delta^{\mathsf{M}}_{S,\varnothing}(c_0) = \langle \varphi_{\mathcal{H}}(S) - \varphi_{\mathcal{H}}(\varnothing), D^{\mathsf{M}}(c, c_0)\rangle.$$

Hence the fixed representation $\varphi(S) = \varphi_{\mathcal{H}}(S)$ satisfies the attribute-potential assumption of Thm. 4.

It remains only to verify that each target vector $\varphi_{\mathcal{H}}(S)$, for $S \subseteq [m]$, belongs to the affine hull of the observed vectors $\{\varphi_{\mathcal{H}}(U) \mid U \in \mathcal{H}\}$. This is a standard consequence of Möbius inversion, equivalently of the invertibility of the zeta matrix of a finite poset, in the incidence-algebra formulation of Rota [1964]. For completeness, we recall the short argument.

We first show linear that the observed vectors $\{\varphi_{\mathcal{H}}(U) \mid U \in \mathcal{H}\}$ are linearly independent. Suppose $\sum_{U\in\mathcal{H}} \alpha_U \varphi_{\mathcal{H}}(U) = 0$. If some coefficient is nonzero, choose an inclusion-maximal $U_0 \in \mathcal{H}$ among the sets with $\alpha_{U_0} \neq 0$, which is possible because $\mathcal{H}$ is finite. Looking at the U_0-coordinate of the vector equality gives

$$0 = \sum_{U\in\mathcal{H}} \alpha_U \varphi_{\mathcal{H}}(U)_{U_0} = \sum_{\substack{U\in\mathcal{H}\\ U_0\subseteq U}} \alpha_U.$$

By maximality of U_0, every term in the last sum except $U = U_0$ has coefficient zero. Therefore $0 = \alpha_{U_0}$, contradicting the choice of U_0. Thus all coefficients are zero, so the observed vectors are linearly independent. There are exactly $|\mathcal{H}|$ such vectors in the $|\mathcal{H}|$-dimensional space $\mathbb{R}^{\mathcal{H}}$, so they form a basis of $\mathbb{R}^{\mathcal{H}}$.

Now fix any target $S \subseteq [m]$. Since the observed vectors form a basis, there exist coefficients $\alpha_U(S)$, $U \in \mathcal{H}$, such that

$$\varphi_{\mathcal{H}}(S) = \sum_{U\in\mathcal{H}} \alpha_U(S) \varphi_{\mathcal{H}}(U).$$

This expansion is automatically affine rather than merely linear. Indeed, looking at the $\varnothing$-coordinate gives

$$1 = \varphi_{\mathcal{H}}(S)_{\varnothing} = \sum_{U\in\mathcal{H}} \alpha_U(S) \varphi_{\mathcal{H}}(U)_{\varnothing} = \sum_{U\in\mathcal{H}} \alpha_U(S),$$

because $\varnothing \subseteq S$ and $\varnothing \subseteq U$ for every $U \in \mathcal{H}$. Therefore $\varphi_{\mathcal{H}}(S)$ lies in the affine hull of $\{\varphi_{\mathcal{H}}(U) \mid U \in \mathcal{H}\}$. Since $S \subseteq [m]$ was arbitrary, $\varphi_{\mathcal{H}}(\mathcal{A}_{\text{ex}})$ is contained in the affine hull of $\varphi_{\mathcal{H}}(\mathcal{A}_{\text{o}})$.

Let $\mathsf{M}, \mathsf{M}' \in \mathcal{M}$ be feature-equivalent on $\mathcal{A}_{\text{o}} = \mathcal{H}$. The preceding two paragraphs verify the hypotheses of Thm. 4 with $\varphi = \varphi_{\mathcal{H}}$. Therefore the transported attribute-potential identities that hold on the observed attributes $\mathcal{H}$ extend to every $S \subseteq [m]$. By Thm. 3, these extended attribute-potential identities imply $\mathbf{P}^{\mathsf{M}}_S = \mathbf{P}^{\mathsf{M}'}_S$ for every $S \subseteq [m]$. Hence M and M' are feature-equivalent on all of $2^{[m]}$. □

D Deferred details

D.1 Details for the running example

We provide the linear-algebra details behind Ex. 2. Fix $\tau \in \mathcal{T}^{\nu}_{\mathbf{Q}}(\bar{\mathcal{Q}}; \mathcal{A}_{\text{o}})$. By Thm. 2, there exists $\mathbf{Q}' = \mathbf{T}_{f'} \in \mathcal{Q}$ such that, after centering at any fixed c_0,

$$\langle f'(a) - f'(a_0), c - c_0\rangle = \langle f(a) - f(a_0), \tau(c) - \tau(c_0)\rangle \qquad \forall a \in \mathcal{A}_{\text{o}}.$$

Let

$$V := \operatorname{span}\{f(a) - f(a_0) \mid a \in \mathcal{A}_{\text{o}}\}.$$

Choose $a_1, \ldots, a_r \in \mathcal{A}_{\text{o}}$ such that $f(a_i) - f(a_0)$, $i = 1, \ldots, r$, form a basis of V. Define

$$F_o := (f(a_1) - f(a_0), \ldots, f(a_r) - f(a_0))^{\top}, \qquad F'_o := (f'(a_1) - f'(a_0), \ldots, f'(a_r) - f'(a_0))^{\top}.$$

Let V' be the row space of F'_o, and let Π_V and $\Pi_{V'}$ denote the orthogonal projections onto V and V', respectively. Stacking the preceding identities gives

$$F'_o(c - c_0) = F_o(\tau(c) - \tau(c_0)).$$

Since F_o has full row rank and row space V,

$$\Pi_V = F_o^\top (F_o F_o^\top)^{-1} F_o.$$

Therefore,

$$\Pi_V \tau(c) = M \Pi_{V'} c + b, \qquad M := F_o^\top (F_o F_o^\top)^{-1} F'_o, \qquad b := \Pi_V \tau(c_0) - M \Pi_{V'} c_0.$$

This proves the projection identity used in the partial-identifiability remark.

If $V = \mathbb{R}^k$, then F_o has full column rank. The stacked identity forces F'_o to have full column rank as well; otherwise $F_o(\tau(c) - \tau(c_0))$ would lie in a proper subspace for all c, contradicting $\tau \in \mathrm{Aut}_\nu(\mathbb{R}^k)$. Hence $V' = \mathbb{R}^k$, $\Pi_V = \Pi_{V'} = I$, and the projection identity reduces to

$$\tau(c) = Mc + b.$$

Thus, under the full-span condition,

$$\mathcal{T}^\nu_{\bar{\mathbf{Q}}}(\bar{\mathcal{Q}}; \mathcal{A}_\mathrm{o}) \subseteq \mathrm{Affine}(k) \cap \mathrm{Aut}_\nu(\mathbb{R}^k).$$

D.2 Details for factorial perturbation modeling

We illustrate Cor. 1 in a factorial experiment with three perturbations. The attribute $S \in 2^{[3]}$ records the applied perturbations, the concept $C \in \mathbb{R}^k$ is the latent state, and the feature X is its observed measurement.

Kernels and assumptions. Fix a positive finite base density q on $\mathbb{R}^k$ and bounded measurable sufficient statistics $T\colon \mathbb{R}^k \to \mathbb{R}^r$. For vectors $\theta_i^{\mathsf{M}}, \theta_{ij}^{\mathsf{M}} \in \mathbb{R}^r$, set

$$\lambda_S^{\mathsf{M}} = \sum_{i \in S} \theta_i^{\mathsf{M}} + \sum_{\{i,j\} \subseteq S} \theta_{ij}^{\mathsf{M}}, \qquad \mathbf{Q}^{\mathsf{M}} = \mathbf{T}_{S \mapsto \lambda_S^{\mathsf{M}}}, \qquad \mathbf{B}(dc \mid \lambda) = \frac{q(c) e^{\langle \lambda, T(c) \rangle}}{Z(\lambda)}\, dc.$$

Here λ_S^{M} parameterizes the interaction energy, with $h_i^{\mathsf{M}}(c) = \langle \theta_i^{\mathsf{M}}, T(c) \rangle$ and $h_{ij}^{\mathsf{M}}(c) = \langle \theta_{ij}^{\mathsf{M}}, T(c) \rangle$. The fixed $\mathbf{B}$ converts this energy into a normalized concept law, while the indexing class restricts $S \mapsto \lambda_S^{\mathsf{M}}$ to main and pairwise terms. Boundedness of T ensures $0 < Z(\lambda) < \infty$ for every λ, so the induced laws have common Lebesgue support.

The observation kernel is $\mathbf{K}^{\mathsf{M}}(\cdot \mid c)$. To apply Thm. 1, its class must be Blackwell reducible. A concrete continuous-measurement example is $X = g(C) + \varepsilon$ with injective g and a shared independent Gaussian noise ε.

Potential and identifiable effects. Let $\mathcal{A}_\mathrm{o} = \{\varnothing, 1, 2, 3, 12, 13, 23\}$, where ij abbreviates $\{i, j\}$. Relative to the control $\varnothing$, define the centered potential

$$\widetilde{\Delta}_S^{\mathsf{M}}(c, c_0) = \Delta_{S,\varnothing}^{\bar{\mathbf{Q}}^{\mathsf{M}}}(c) - \Delta_{S,\varnothing}^{\bar{\mathbf{Q}}^{\mathsf{M}}}(c_0) = \sum_{i \in S} \delta h_i^{\mathsf{M}}(c, c_0) + \sum_{\{i,j\} \subseteq S} \delta h_{ij}^{\mathsf{M}}(c, c_0),$$

where $\delta h_T^{\mathsf{M}}(c, c_0) = h_T^{\mathsf{M}}(c) - h_T^{\mathsf{M}}(c_0)$. If M, M' are feature-equivalent on $\mathcal{A}_\mathrm{o}$, Thms. 1 and 2 give a single transition τ with

$$\widetilde{\Delta}_S^{\mathsf{M}'}(c, c_0) = \widetilde{\Delta}_S^{\mathsf{M}}(\tau(c), \tau(c_0)), \qquad S \in \mathcal{A}_\mathrm{o},$$

for Lebesgue-almost every pair (c, c_0). Since $\delta h_i^{\mathsf{M}} = \widetilde{\Delta}_i^{\mathsf{M}}$ and $\delta h_{ij}^{\mathsf{M}} = \widetilde{\Delta}_{ij}^{\mathsf{M}} - \widetilde{\Delta}_i^{\mathsf{M}} - \widetilde{\Delta}_j^{\mathsf{M}}$, the centered main and pair effects obey the same transport identity. Thus these effects are identified up to additive constants and the common latent change of coordinates allowed by the CMM class.

Extrapolation to the triple. Counting each main and pair term gives, in each candidate model,

$$\widetilde{\Delta}_{123}^{\mathsf{M}} = \widetilde{\Delta}_{12}^{\mathsf{M}} + \widetilde{\Delta}_{13}^{\mathsf{M}} + \widetilde{\Delta}_{23}^{\mathsf{M}} - \widetilde{\Delta}_1^{\mathsf{M}} - \widetilde{\Delta}_2^{\mathsf{M}} - \widetilde{\Delta}_3^{\mathsf{M}}.$$

Transporting the six observed terms therefore transports the centered triple potential. The anchor density identity and normalization remove the remaining additive constant, as in the proof of Thm. 4. Thm. 3 yields equality of the full feature distributions at the held-out triple.

Valid transitions and failure cases. A valid transition must satisfy both $\mathbf{K}' \overset{\nu}{=} \mathbf{K}\mathbf{T}_\tau$ and the transported concept-law identities, including the control anchor. For an explicit valid transition, take q to be the standard Gaussian density, $T(c) = (\tanh c_i)_{i=1}^k$, and let P be a coordinate permutation matrix. Then $\tau(c) = Pc$ is valid with $\theta_T^{\mathsf{M}'} = P^\top \theta_T^{\mathsf{M}}$ for every main or pair term T, and $g' = g \circ P$ in the continuous observation example. Indeed, $q(Pc) = q(c)$ and $T(Pc) = PT(c)$, so these choices transport all concept laws and preserve the feature laws. Scaling, for instance by 2, is invalid in this same class, because $\lambda_\varnothing = 0$ fixes the anchor law to q and scaling changes its covariance.

The interaction restriction is essential for the extrapolation claim. If the class permits an additional bounded, nonconstant term $h_{123}(c)\mathbf{1}\{S = 123\}$, one may change that term while preserving every observed law and the control. The triple concept law changes; under an injective observation channel its feature law changes as well. This gives a feature-equivalent alternative on $\mathcal{A}_{\mathrm{o}}$ with a different prediction at the target.

E Recoveries of representative prior work

The recoveries below are not intended as new proofs of the cited identifiability theorems. Instead, each recovery is organized as a short CMM dictionary followed by the CMM proof object that the original paper analyzes. The generic transition step is supplied by Thm. 1 and, when a density-level statement is needed, by Thm. 2. The remaining rigidity step is delegated to the corresponding theorem or proof in the cited work.

Shared architecture of the recoveries. Each recovery follows the same template:

(i) **CMM translation.** Each recovery first translates the prior model into CMM notation by specifying the attribute A, the modulator Λ, the concept C, the feature X, and the kernels $\mathbf{Q}$, $\mathbf{B}$, and $\mathbf{K}$.

(ii) **Attribute potentials** or **Proof object.** The next paragraph computes the attribute potential or the equivalent object used by the cited proof, such as a centered potential, exponentiated density ratio, score contrast, covariance equation, or quadratic transition identity.

(iii) **Recovery of guarantees.** The proposition starts from feature-equivalent CMMs, applies the CMM transition theorem, rewrites the transported identity as the cited paper's proof object, and invokes the paper-specific rigidity result that yields the final ambiguity class.

Accordingly, the short proofs below only verify the CMM-to-paper identity displayed in the proposition; they do not reprove the paper-specific rigidity arguments.

Applicability of Thm. 1. In the continuous recoveries below, we take ν to be Lebesgue measure on the corresponding concept space $\mathcal{C} = \mathbb{R}^d$ or $\mathbb{R}^n$. Except where explicitly noted, the mixing kernels are deterministic kernels $\mathbf{T}_g$, where $g\colon \mathcal{C} \to \mathcal{X}$ is an admissible diffeomorphic embedding; in linear CRL, g is the full-column-rank linear mixing matrix G, which is a bimeasurable embedding of $\mathbb{R}^d$ onto its image. Thus the mixing classes are ν-Blackwell reducible by taking $\widetilde{\mathcal{X}} = \mathcal{X}$, $\widetilde{\mathbf{K}} = \mathbf{T}_{\mathrm{id}_\mathcal{X}}$, and $\mathcal{G}$ to be the relevant class of admissible embeddings, since $\xi \mapsto \widetilde{\mathbf{K}}\xi$ is the identity map on the pushed-forward measures. For noisy iVAE-type observation models, write $\mathbf{K} = \widetilde{\mathbf{K}}\mathbf{T}_g$, where g is the decoder embedding and $\widetilde{\mathbf{K}}$ is the shared observation-noise channel. For independent additive noise with a nowhere-zero characteristic function, including Gaussian noise, this channel is injective on probability laws: equality after convolution implies equality of characteristic functions before convolution. This verifies Blackwell reducibility for the noisy model directly, without preprocessing or a deconvolution estimator.

On the concept side, we assume the common-carrier condition required by Thm. 1: the induced concept laws in the model class are dominated by ν, and their densities are positive and finite ν-a.e. on the attributes under consideration. This condition holds directly for the Gaussian CRL, score-based CRL, and nonparametric CRL recoveries under their full-support density assumptions, and for nonlinear ICA / iVAE under the usual positivity assumptions on the conditional source or prior densities. For linear CRL [Squires et al., 2023], the original theorem is covariance-based and does not require densities; when we invoke Thm. 1, we are considering the subclass satisfying the additional common-carrier condition, for example when the noise ϵ has a strictly positive finite Lebesgue density

and all B_k are invertible. Under these standing conditions, any feature-equivalent alternative admits the transition supplied by Thm. 1; the individual recoveries below then identify the corresponding proof object and delegate the model-specific rigidity step to the cited paper.

E.1 Nonlinear ICA

CMM translation. For nonlinear ICA [Hyvarinen et al., 2019], take A to be the auxiliary variable, $\mathcal{C} = \mathbb{R}^k$, $C = Z$, and $X = f(Z)$. For each attribute value a, define the modulator value

$$\lambda_a := \big(q_1(\cdot, a), \dots, q_k(\cdot, a)\big).$$

Thus the modulator contains only the coordinate-wise conditional source log-potentials. The normalizer is the derived functional

$$\Gamma(\lambda) := \log \int_{\mathbb{R}^k} \exp\left\{\sum_{i=1}^{k} \lambda_i(c_i)\right\} dc,$$

and we write $\Gamma(a) := \Gamma(\lambda_a)$. The deterministic indexing kernel is

$$\mathbf{Q} = \mathbf{T}_{a \mapsto \lambda_a}.$$

The shared concept-modulation kernel maps a tuple of source log-potentials to the corresponding conditionally independent source density:

$$\mathbf{B}(dc \mid \lambda) = \exp\left\{\sum_{i=1}^{k} \lambda_i(c_i) - \Gamma(\lambda)\right\} dc.$$

Thus, for $\bar{\mathbf{Q}} = \mathbf{B}\mathbf{Q}$,

$$\log p_{\bar{\mathbf{Q}}}(c \mid a) = \sum_{i=1}^{k} q_i(c_i, a) - \Gamma(a).$$

The mixing kernel is deterministic, $\mathbf{K} = \mathbf{T}_f$.

Attribute potentials. Relative to an anchor a_0, the attribute potential is

$$\Delta^{\bar{\mathbf{Q}}}_{a,a_0}(c) = \sum_{i=1}^{k} \{q_i(c_i, a) - q_i(c_i, a_0)\} - \{\Gamma(a) - \Gamma(a_0)\}.$$

The normalizer contrast is independent of c, so all mixed second derivatives of the potential are determined only by the coordinate-wise contrasts

$$h_i(u; a, a_0) := q_i(u, a) - q_i(u, a_0).$$

This separable potential is the CMM proof object whose transported mixed derivatives reproduce the nonlinear ICA equations of Hyvarinen et al. [2019].

Recovery of guarantees.

Proposition 2 (CMM recovery of Theorem 1 of [Hyvarinen et al., 2019])**.** *Let*

$$\mathsf{M} = (\mathbf{T}_{a \mapsto \lambda_a}, \mathbf{B}, \mathbf{T}_f), \qquad \mathsf{M}' = (\mathbf{T}_{a \mapsto \lambda'_a}, \mathbf{B}, \mathbf{T}_{f'})$$

be feature-equivalent nonlinear ICA CMMs satisfying the common-support and smoothness assumptions needed for Thms. 1 and 2. Let $\tau = f^{-1} \circ f'$ be the transition induced by Thm. 1. For $h_i(u; a, a_0) := q_i(u, a) - q_i(u, a_0)$, the transported CMM potential identity implies, for every $j \neq j'$,

$$0 = \sum_{i=1}^{k} \partial_1 h_i(\tau_i(c); a, a_0) \partial^2_{c_j c_{j'}} \tau_i(c) + \sum_{i=1}^{k} \partial^2_1 h_i(\tau_i(c); a, a_0) \partial_{c_j} \tau_i(c) \partial_{c_{j'}} \tau_i(c).$$

This is the mixed-derivative system used in the proof of Theorem 1 of Hyvarinen et al. [2019]. Consequently, under their Assumption of Variability, τ is componentwise up to permutation:

$$\tau(c) = \big(\phi_1(c_{\pi(1)}), \dots, \phi_k(c_{\pi(k)})\big).$$

Proof. By Thm. 2, feature equivalence gives

$$\Delta^{\bar{\mathbf{Q}}'}_{a,a_0}(c) = \Delta^{\bar{\mathbf{Q}}}_{a,a_0}(\tau(c)).$$

The candidate potential $\Delta^{\bar{\mathbf{Q}}'}_{a,a_0}(c)$ is coordinate-separable in c up to an additive constant, so

$$\partial^2_{c_j c_{j'}} \Delta^{\bar{\mathbf{Q}}'}_{a,a_0}(c) = 0 \qquad \text{for every } j \neq j'.$$

Applying the same mixed derivative to the transported right-hand side gives

$$0 = \partial^2_{c_j c_{j'}} \left[\sum_{i=1}^{k} h_i(\tau_i(c); a, a_0) \right],$$

because $\Gamma(a) - \Gamma(a_0)$ is independent of c. Expanding this derivative by the chain rule gives the displayed mixed-derivative system. The remaining step is exactly the variability-rank and integration argument in Hyvarinen et al. [2019], which forces a monomial Jacobian and hence componentwise recovery up to permutation. □

E.2 Conditionally exponential families

CMM translation. For the conditionally exponential-family model including iVAE [Khemakhem et al., 2020a] and ICE-BeeM [Khemakhem et al., 2020b], the concept density has the form

$$p_{\bar{\mathbf{Q}}}(c \mid a) = \frac{Q(c)}{Z(a)} \exp\{\langle f(a), T(c)\rangle\}, \qquad X = g(C),$$

where $f\colon \mathcal{A} \to \mathbb{R}^m$ is the attribute-dependent natural parameter, $T\colon \mathcal{C} \to \mathbb{R}^m$ collects sufficient statistics, Q is a positive base density, and Z is the normalizer. The modulator is $\lambda(a) = (f(a), T, Q)$, with indexing kernel $\mathbf{Q} = \mathbf{T}_\eta$ for $\eta(a) = \lambda(a)$. The shared concept-modulation kernel maps (θ, T, Q) to the density proportional to $Q(c)\exp\{\langle \theta, T(c)\rangle\}$. The mixing kernel is $\mathbf{K} = \mathbf{T}_g$, or $\mathbf{K} = \widetilde{\mathbf{K}}\mathbf{T}_g$ in the noisy-observation version.

Attribute potentials. The attribute potential relative to a_0 is

$$\Delta^{\bar{\mathbf{Q}}}_{a,a_0}(c) = \langle f(a) - f(a_0), T(c)\rangle - \{\log Z(a) - \log Z(a_0)\}.$$

After centering at any c_0, the normalizers cancel:

$$\Delta^{\bar{\mathbf{Q}}}_{a,a_0}(c) - \Delta^{\bar{\mathbf{Q}}}_{a,a_0}(c_0) = \langle f(a) - f(a_0), T(c) - T(c_0)\rangle.$$

This centered potential is the affine sufficient-statistic equation used in the iVAE identifiability proof.

Recovery of guarantees.

Proposition 3 (CMM recovery of Theorem 1 of [Khemakhem et al., 2020a])**.** *Let* $\mathsf{M} = (\mathbf{T}_\eta, \mathbf{B}, \mathbf{T}_g)$ *and* $\mathsf{M}' = (\mathbf{T}_{\eta'}, \mathbf{B}, \mathbf{T}_{g'})$ *be feature-equivalent conditionally exponential-family CMMs, with* $\eta(a) = (f(a), T, Q)$ *and* $\eta'(a) = (f'(a), T', Q')$*. Let* $\tau = g^{-1} \circ g'$ *be the transition induced by Thm. 1. Assume there are attributes* $a_0, a_1, \ldots, a_m$ *such that*

$$L := \big(f(a_1) - f(a_0), \ldots, f(a_m) - f(a_0)\big)$$

is invertible, and define L' *analogously. Then the CMM potential identity gives*

$$T(\tau(c)) = MT'(c) + b, \qquad M = L^{-\top}(L')^\top,$$

for some $b \in \mathbb{R}^m$*. This is the* $\sim_A$*-identifiability relation of Theorem 1 of Khemakhem et al. [2020a]. The further refinement to* $\sim_P$ *is the separate rigidity step in their Theorems 2–3.*

Proof. By Thm. 2, $\Delta^{\bar{\mathbf{Q}}'}_{a,a_0}(c) = \Delta^{\bar{\mathbf{Q}}}_{a,a_0}(\tau(c))$. Centering at c_0 cancels the normalizers and gives

$$\langle f'(a) - f'(a_0), T'(c) - T'(c_0)\rangle = \langle f(a) - f(a_0), T(\tau(c)) - T(\tau(c_0))\rangle.$$

Stacking this identity over $a_1, \ldots, a_m$ and solving with L yields

$$T(\tau(c)) - T(\tau(c_0)) = L^{-\top}(L')^\top\{T'(c) - T'(c_0)\}.$$

Absorbing the value at c_0 into b gives the affine relation. □

E.3 Linear CRL

CMM translation. For the linear CRL model of Squires et al. [2023], let $\mathcal{A} = \{0\} \cup [K]$, $\mathcal{C} = \mathbb{R}^d$, and $\mathcal{X} = \mathbb{R}^p$. The attribute $k \in \mathcal{A}$ indexes an observational or interventional environment, the concept is $C = Z^{(k)}$, and the feature is $X^{(k)} = GZ^{(k)}$ with $G \in \mathbb{R}^{p\times d}$ full column rank. In environment k, the latent variables satisfy

$$Z^{(k)} = B_k^{-1}\epsilon, \qquad \mathbb{E}[\epsilon] = 0, \qquad \mathrm{Cov}(\epsilon) = I_d,$$

where $B_k = \Omega_k^{-1/2}(I_d - A_k)$ is the normalized structural factor obtained through a perfect intervention: doing such on a node i_k changes only the i_k-th row of B_0, replacing it by $\lambda_k e_{i_k}^\top$. The indexing kernel is $\mathbf{Q} = \mathbf{T}_\eta$, where $\eta(k) = B_k$. The concept-modulation kernel maps B to the law of $B^{-1}\epsilon$, and the mixing kernel is deterministic, $\mathbf{K} = \mathbf{T}_G$.

Second-moment proof object. Writing $H = G^\dagger$, the observed pseudoprecision in environment k is

$$\Theta_k := \mathrm{Cov}(X^{(k)})^\dagger = H^\top B_k^\top B_k H.$$

The residual relabeling set is

$$S(\mathcal{G}) := \{\sigma\colon [d] \to [d] \text{ bijective} \mid \sigma(j) > \sigma(i) \text{ for every edge } j \to i \text{ in } \mathcal{G}\}.$$

Thus the proof object in Squires et al. [2023] is not an attribute-potential equation but the family of observed pseudoprecision equations.

Recovery of guarantees.

Proposition 4 (CMM recovery of Theorem 2 of [Squires et al., 2023])**.** *Let* $\mathsf{M} = (\mathbf{T}_\eta, \mathbf{B}, \mathbf{T}_G)$ *and* $\mathsf{M}' = (\mathbf{T}_{\widetilde{\eta}}, \mathbf{B}, \mathbf{T}_{\widetilde{G}})$ *be feature-equivalent linear CRL CMMs with full-column-rank mixing matrices. Then the mixing side of Thm. 1 forces the induced transition to be linear:*

$$\widetilde{G} = GT, \qquad T = G^\dagger \widetilde{G} \in \mathrm{GL}(d).$$

The concept-side equality gives

$$\mathrm{Cov}(Z^{(k)}) = T\,\mathrm{Cov}(\widetilde{Z}^{(k)})\,T^\top,$$

and hence

$$\Theta_k = H^\top B_k^\top B_k H = \widetilde{H}^\top \widetilde{B}_k^\top \widetilde{B}_k \widetilde{H}.$$

Under Assumptions 1–2 of Squires et al. [2023] with one perfect intervention per latent node, their Theorem 2 gives identifiability up to $S(\mathcal{G})$.

Proof. From the mixing side relation $\mathbf{T}_{\widetilde{G}} \overset{\nu}{=} \mathbf{T}_G \mathbf{T}_\tau$, we have $\widetilde{G}z = G\tau(z)$ for ν-a.e. z. Multiplying by $G^\dagger$ gives $\tau(z) = Tz$ with $T = G^\dagger\widetilde{G}$, and full column rank makes T invertible. The concept-side equality transports covariances by T, which is equivalent to the displayed pseudoprecision equations. These are exactly the equations analyzed in Squires et al. [2023]; their generalized-RQ and partial-order argument gives the stated residual ambiguity. □

Remark 2. *This recovery is intentionally mixing-side rather than potential-ratio based. Unlike the nonlinear recoveries, the restrictive object in linear CRL is the linear mixing kernel* $\mathbf{K} = \mathbf{T}_G$*. The CMM transition is forced to be linear before applying the model-specific rigidity theorem of Squires et al. [2023].*

E.4 Gaussian CRL

CMM translation. For the Gaussian CRL model of Buchholz et al. [2023], let $\mathcal{A} = I \cup \{0\}$, $\mathcal{C} = \mathbb{R}^d$, and $\mathcal{X} = \mathbb{R}^{d'}$. The attribute $i \in I$ indexes an interventional environment, while 0 is the observational environment. The latent concept is $C = Z^{(i)}$, and the observed feature is $X^{(i)} = f(Z^{(i)})$. In environment i, the structural factor is

$$B^{(i)} = (D^{(i)})^{-1/2}(I_d - A^{(i)}),$$

and only the row indexed by the target node t_i differs from the observational factor $B^{(0)}$. The scalar $\eta^{(i)}$ denotes the intervention shift parameter from Buchholz et al. [2023]; it is not an indexing map.

Define the environment-specific modulator value by

$$\lambda_i := (B^{(i)}, \eta^{(i)}, t_i), \qquad i \in I \cup \{0\},$$

with $\eta^{(0)} = 0$ and arbitrary t_0, since the shift term vanishes in the observational environment. The deterministic indexing kernel is

$$\mathbf{Q} = \mathbf{T}_{i \mapsto \lambda_i}.$$

The shared concept-modulation kernel sends a structural triple (B, α, t) to the Gaussian law induced by $B^{-1}(\epsilon + \alpha e_t)$:

$$\mathbf{B}(dc \mid B, \alpha, t) = |\det B|\,(2\pi)^{-d/2} \exp\left\{-\frac{1}{2}\|Bc - \alpha e_t\|^2\right\} dc.$$

Thus, for $\bar{\mathbf{Q}} = \mathbf{B}\mathbf{Q}$,

$$p_{\bar{\mathbf{Q}}}(c \mid i) = |\det B^{(i)}|\,(2\pi)^{-d/2} \exp\left\{-\frac{1}{2}\|B^{(i)}c - \eta^{(i)} e_{t_i}\|^2\right\}.$$

Equivalently,

$$C \mid A = i \sim Z^{(i)} = (B^{(i)})^{-1}(\epsilon + \eta^{(i)} e_{t_i}), \qquad \epsilon \sim \mathcal{N}(0, I_d).$$

The mixing kernel is deterministic, $\mathbf{K} = \mathbf{T}_f$.

Attribute potentials. Write

$$r^{(i)} := (B^{(i)})^\top e_{t_i}, \qquad \Theta^{(i)} := (B^{(i)})^\top B^{(i)}.$$

The attribute potential relative to 0 is

$$\Delta^{\bar{\mathbf{Q}}}_{i,0}(c) = -\frac{1}{2} c^\top (\Theta^{(i)} - \Theta^{(0)}) c + \eta^{(i)} (r^{(i)})^\top c + \kappa^{(i)},$$

where

$$\kappa^{(i)} = \log \frac{|\det B^{(i)}|}{|\det B^{(0)}|} - \frac{1}{2}(\eta^{(i)})^2$$

is independent of c. Equivalently, up to an i-dependent additive constant, $-\Delta^{\bar{\mathbf{Q}}}_{i,0}$ is the quadratic-linear form

$$\frac{1}{2} c^\top (\Theta^{(i)} - \Theta^{(0)}) c - \eta^{(i)} (r^{(i)})^\top c.$$

This is the quadratic transition object used in Buchholz et al. [2023].

Recovery of guarantees.

Proposition 5 (CMM recovery of Theorems 1 and 3 of [Buchholz et al., 2023])**.** *Let* M *and* M' *be feature-equivalent Gaussian CRL CMMs. Write the candidate quantities with tildes, so that the candidate mixing map is* $\widetilde{f}$*, the candidate structural factors are* $\widetilde{B}^{(i)}$*, the candidate shift parameters are* $\widetilde{\eta}^{(i)}$*, and*

$$\widetilde{r}^{(i)} := (\widetilde{B}^{(i)})^\top e_{\widetilde{t}_i}, \qquad \widetilde{\Theta}^{(i)} := (\widetilde{B}^{(i)})^\top \widetilde{B}^{(i)}.$$

Let

$$\tau := \widetilde{f}^{-1} \circ f$$

be the transition from ground-truth latent coordinates to candidate latent coordinates. Then, for every intervention $i \in I$*, there exists* $b^{(i)} \in \mathbb{R}$ *such that*

$$\frac{1}{2} c^\top (\Theta^{(i)} - \Theta^{(0)}) c - \eta^{(i)} (r^{(i)})^\top c = \frac{1}{2} \tau(c)^\top (\widetilde{\Theta}^{(i)} - \widetilde{\Theta}^{(0)}) \tau(c) - \widetilde{\eta}^{(i)} (\widetilde{r}^{(i)})^\top \tau(c) + b^{(i)}.$$

This is the quadratic transition identity used in the proof of Theorem 3 of Buchholz et al. [2023]. Under the assumptions of their Theorem 3, their rigidity argument implies $\tau(c) = Tc$*, equivalently* $\widetilde{f} = f \circ T^{-1}$*. Under the additional perfect-intervention assumptions of their Theorem 1, their Appendix B combines this linearity with Squires et al. [2023] to obtain permutation-and-scaling identifiability.*

Proof. Feature equivalence and Thm. 1 give a latent transition. Using the inverse-direction transition $\tau = \widetilde{f}^{-1} \circ f$, Thm. 2 gives

$$\Delta^{\bar{\mathbf{Q}}}_{i,0}(c) = \Delta^{\bar{\mathbf{Q}}'}_{i,0}(\tau(c)).$$

Substituting the Gaussian potential expansions for $\bar{\mathbf{Q}}$ and $\bar{\mathbf{Q}}'$ gives

$$-\frac{1}{2}c^\top(\Theta^{(i)} - \Theta^{(0)})c + \eta^{(i)}(r^{(i)})^\top c + \kappa^{(i)} = -\frac{1}{2}\tau(c)^\top(\widetilde{\Theta}^{(i)} - \widetilde{\Theta}^{(0)})\tau(c) + \widetilde{\eta}^{(i)}(\widetilde{r}^{(i)})^\top\tau(c) + \widetilde{\kappa}^{(i)}.$$

Multiplying by -1 and absorbing the constant $\widetilde{\kappa}^{(i)} - \kappa^{(i)}$ into $b^{(i)}$ gives the displayed quadratic identity. □

E.5 Score-based CRL

CMM translation. In the setting of Varıcı et al. [2025], take $\mathcal{A} = \mathcal{E}, \mathcal{C} = \mathbb{R}^n$, and let the modulator space be the space of admissible local-mechanism values

$$\mathscr{L} = \bigsqcup_{G \in \mathrm{DAG}([n])} \prod_{i=1}^{n} \mathcal{K}^{\mathrm{mk}}(\mathcal{C}_{\mathrm{pa}_G(i)} \to \mathcal{C}_i).$$

Each $\lambda \in \mathscr{L}$ specifies local conditional densities $(p_i^\lambda)_{i=1}^n$ and induces a DAG $\mathcal{G}_\lambda$ on $[n]$. For $i \in [n]$, write $\mathrm{pa}_{\mathcal{G}_\lambda}(i)$ for the parent set induced by λ. Let $\eta\colon \mathcal{E} \to \mathscr{L}$ be the deterministic indexing map and write $\lambda_a := \eta(a)$. The indexing kernel is $\mathbf{Q} = \mathbf{T}_\eta$. The shared concept-modulation kernel maps a mechanism value to the Markov density over its induced graph:

$$\mathbf{B}(dc \mid \lambda) = \prod_{i=1}^{n} p_i^\lambda(c_i \mid c_{\mathrm{pa}_{\mathcal{G}_\lambda}(i)})\, dc.$$

Thus, for $\bar{\mathbf{Q}} = \mathbf{B}\mathbf{T}_\eta$,

$$p_{\bar{\mathbf{Q}}}(c \mid a) = \prod_{i=1}^{n} p_i^{\lambda_a}(c_i \mid c_{\mathrm{pa}_{\mathcal{G}_{\lambda_a}}(i)}).$$

The mixing kernel is deterministic, $\mathbf{K} = \mathbf{T}_f$.

Attribute potentials. For $a, a' \in \mathcal{E}$,

$$\Delta^{\bar{\mathbf{Q}}}_{a,a'}(c) = \sum_{i=1}^{n} \log p_i^{\lambda_a}(c_i \mid c_{\mathrm{pa}_{\mathcal{G}_{\lambda_a}}(i)}) - \sum_{i=1}^{n} \log p_i^{\lambda_{a'}}(c_i \mid c_{\mathrm{pa}_{\mathcal{G}_{\lambda_{a'}}}(i)}).$$

For a coupled hard-intervention pair $(a_i, \tilde{a}_i)$ targeting node i, the two mechanism values agree at all non-i local mechanisms and have parent-free i-th mechanisms. Hence

$$\Delta^{\bar{\mathbf{Q}}}_{a_i,\tilde{a}_i}(c) = \log p_i^{\lambda_{a_i}}(c_i) - \log p_i^{\lambda_{\tilde{a}_i}}(c_i),$$

so $\nabla_c \Delta^{\bar{\mathbf{Q}}}_{a_i,\tilde{a}_i}(c)$ is one-sparse. The differentiated attribute potential is the latent score-difference object used in score-based CRL.

Recovery of guarantees.

Proposition 6 (CMM recovery of Theorem 25 of [Varıcı et al., 2025])**.** *Let* $\mathsf{M} = (\mathbf{T}_\eta, \mathbf{B}, \mathbf{T}_f)$ *and* $\mathsf{M}' = (\mathbf{T}_{\eta'}, \mathbf{B}, \mathbf{T}_{f'})$ *be feature-equivalent score-based CRL CMMs. Write* $\lambda_a := \eta(a)$ *and* $\lambda'_a := \eta'(a)$*, so that the graphs in environment* a *are* $\mathcal{G}_{\lambda_a}$ *and* $\mathcal{G}_{\lambda'_a}$*. Let* τ *be the transition induced by Thm. 1. For any coupled hard-intervention pair* $(a_i, \tilde{a}_i)$*,*

$$\nabla_c \Delta^{\bar{\mathbf{Q}}'}_{a_i,\tilde{a}_i}(c) = \partial_c \tau(c)^\top \nabla_c \Delta^{\bar{\mathbf{Q}}}_{a_i,\tilde{a}_i}(\tau(c)).$$

Moreover, the two score contrasts are one-sparse in the corresponding candidate and ground-truth intervention targets. This is the transported score-contrast and sparsity object used by Varıcı et al. [2025]; their Theorem 25 gives componentwise recovery and recovery of the mechanism-induced anchor graph $\mathcal{G}_{\lambda_{a_0}}$ *up to the corresponding permutation of* $\mathcal{G}_{\lambda'_{a_0}}$*.*

Proof. By Thm. 2, subtracting the transported identities for a_i and $\tilde{a}_i$ gives $\Delta^{\bar{\mathbf{Q}}'}_{a_i,\tilde{a}_i}(c) = \Delta^{\bar{\mathbf{Q}}}_{a_i,\tilde{a}_i}(\tau(c))$. Differentiating gives the displayed score-contrast equation. For a coupled hard-intervention pair, all non-target mechanisms cancel and the target mechanisms are parent-free, so the ground-truth and candidate score contrasts are one-sparse in their respective targets. This is exactly the sparsity property formalized in Theorem 7(iii) of Varıcı et al. [2025]; the componentwise and graph-recovery steps are their Lemmas 23–24 and Theorem 25. □

E.6 Nonparametric CRL

CMM translation. For the nonparametric CRL model of von Kügelgen et al. [2023], let $\mathcal{A} = \mathcal{E}$, $\mathcal{C} = \mathbb{R}^n$, and $\mathcal{X} \subseteq \mathbb{R}^d$. Fix an observational anchor $e_0 \in \mathcal{E}$. The concept variable is $C = (C_1, \dots, C_n)$, and the observed variable is $X = f(C)$, where $f\colon \mathbb{R}^n \to \mathcal{X}$ is a diffeomorphism onto its image. Let $\mathscr{L}$ be the space of admissible local-mechanism values. Each $\lambda \in \mathscr{L}$ specifies local conditional densities $(p_j^\lambda)_{j=1}^n$ and induces a DAG $\mathcal{G}_\lambda$ on $[n]$. Let $\eta\colon \mathcal{E} \to \mathscr{L}$ be the deterministic indexing map, write $\lambda_e := \eta(e)$, and set $\lambda_0 := \lambda_{e_0}$. The indexing kernel is $\mathbf{Q} = \mathbf{T}_\eta$. The shared concept-modulation kernel is

$$\mathbf{B}(dc \mid \lambda) = \prod_{j=1}^n p_j^\lambda(c_j \mid c_{\mathrm{pa}_{\mathcal{G}_\lambda}(j)})\, dc.$$

Thus, for $\bar{\mathbf{Q}} = \mathbf{B}\mathbf{T}_\eta$,

$$p_{\bar{\mathbf{Q}}}(c \mid e) = \prod_{j=1}^n p_j^{\lambda_e}(c_j \mid c_{\mathrm{pa}_{\mathcal{G}_{\lambda_e}}(j)}).$$

The mixing kernel is deterministic, $\mathbf{K} = \mathbf{T}_f$, and the ground-truth anchor graph is the mechanism-induced graph $\mathcal{G}_{\lambda_0}$.

Attribute potentials. For $e, e' \in \mathcal{E}$, define

$$R^{\bar{\mathbf{Q}}}_{e,e'}(c) := \exp\{\Delta^{\bar{\mathbf{Q}}}_{e,e'}(c)\} = \frac{p_{\bar{\mathbf{Q}}}(c \mid e)}{p_{\bar{\mathbf{Q}}}(c \mid e')}.$$

For a paired perfect intervention (e, e') on node i, all non-target mechanisms agree and cancel in the ratio. The target mechanisms are parent-free, so with

$$\rho_i^{e,e'}(u) := \frac{\widetilde{p}_i^{e}(u)}{\widetilde{p}_i^{e'}(u)},$$

we have

$$R^{\bar{\mathbf{Q}}}_{e,e'}(c) = \rho_i^{e,e'}(c_i).$$

Thus the CMM proof object is the paired environment density ratio used by von Kügelgen et al. [2023].

Recovery of guarantees.

Proposition 7 (CMM recovery of Theorem 3.4 of [von Kügelgen et al., 2023])**.** *Let* $\mathsf{M} = (\mathbf{T}_\eta, \mathbf{B}, \mathbf{T}_f)$ *and* $\mathsf{M}' = (\mathbf{T}_{\eta'}, \mathbf{B}, \mathbf{T}_{h^{-1}})$ *be feature-equivalent nonparametric CRL CMMs. Write* $\lambda_e := \eta(e)$, $\lambda'_e := \eta'(e)$, $\lambda_0 := \lambda_{e_0}$, *and* $\lambda'_0 := \lambda'_{e_0}$. *Let* $h\colon \mathcal{X} \to \mathbb{R}^n$ *be the candidate unmixing map, and define the transition from candidate to ground-truth coordinates by*

$$\tau(z) := f^{-1}(h^{-1}(z)).$$

For any paired perfect intervention (e, e') *with ground-truth target* i *and candidate target* j,

$$\rho'^{e,e'}_j(z_j) = \rho_i^{e,e'}(\tau_i(z)).$$

For the n *paired interventions used in Theorem 3.4 of von Kügelgen et al. [2023], writing* $j = \pi(i)$ *gives*

$$\rho'^{e,e'}_{\pi(i)}(z_{\pi(i)}) = \rho_i^{e,e'}(\tau_i(z)).$$

This is the one-dimensional ratio equation in Appendix C.3 of von Kügelgen et al. [2023]; their theorem gives coordinatewise recovery and graph isomorphism between the mechanism-induced anchor graphs $\mathcal{G}_{\lambda_0}$ *and* $\mathcal{G}_{\lambda'_0}$.

Proof. By Thm. 2, feature equivalence gives $R^{\bar{\mathbf{Q}}'}_{e,e'}(z) = R^{\bar{\mathbf{Q}}}_{e,e'}(\tau(z))$. For a paired perfect intervention, all non-target mechanisms cancel in both ratios. Thus the ground-truth ratio is $\rho_i^{e,e'}(\tau_i(z))$, while the candidate ratio is $\rho_j'^{e,e'}(z_j)$. Substitution gives the displayed equation. The nondegeneracy, coordinatewise-recovery, and graph-isomorphism steps are exactly those in Appendix C.3 and Theorem 3.4 of von Kügelgen et al. [2023]. □

E.7 Perturbation modeling

CMM translation. Perturbation modeling in von Kügelgen et al. [2025] fits the CMM template by taking the attribute $a \in \mathcal{A} = \mathbb{R}^K$ to be a perturbation label, the concept $c \in \mathcal{C} = \mathbb{R}^k$ to be the perturbation-relevant latent state, and the feature $x \in \mathcal{X}$ to be the observed measurement. Fix an anchor perturbation a_0. The deterministic indexing kernel is

$$\mathbf{Q} = \mathbf{T}_{a \mapsto W(a-a_0)}.$$

The shared concept-modulation kernel is

$$\mathbf{B}(dc \mid \lambda) = \mathcal{N}(\lambda, I_k)(dc)$$

where $\mathcal{N}(\mu, \Sigma)$ denotes a Gaussian distribution with mean μ and covariance Σ. Thus, for $\bar{\mathbf{Q}} = \mathbf{B}\mathbf{Q}$,

$$p_{\bar{\mathbf{Q}}}(c \mid a) = \frac{1}{(2\pi)^{k/2}} \exp\left\{-\frac{1}{2}\|c - W(a-a_0)\|^2\right\}.$$

The mixing kernel is deterministic, $\mathbf{K} = \mathbf{T}_g$, where $g \colon \mathbb{R}^k \to \mathcal{X}$ is the observation map.

Attribute potentials. For this Gaussian mean-shift concept law, the attribute potential relative to a_0 is

$$\begin{aligned}\Delta^{\bar{\mathbf{Q}}}_{a,a_0}(c) &= \log p_{\bar{\mathbf{Q}}}(c \mid a) - \log p_{\bar{\mathbf{Q}}}(c \mid a_0) \\ &= \langle W(a-a_0), c\rangle - \frac{1}{2}\|W(a-a_0)\|^2.\end{aligned}$$

Therefore, after centering at any $c_0 \in \mathbb{R}^k$,

$$\Delta^{\bar{\mathbf{Q}}}_{a,a_0}(c) - \Delta^{\bar{\mathbf{Q}}}_{a,a_0}(c_0) = \langle a - a_0, W^\top(c - c_0)\rangle.$$

This is the proof object used in the perturbation model: perturbations act linearly on the latent mean, and centered attribute potentials expose the perturbation effect matrix W.

Recovery of guarantees.

Proposition 8 (CMM recovery of perturbation identifiability and extrapolation)**.** *Let*

$$\mathsf{M} = (\mathbf{T}_{a \mapsto W(a-a_0)}, \mathbf{B}, \mathbf{T}_g), \qquad \mathsf{M}' = (\mathbf{T}_{a \mapsto W'(a-a_0)}, \mathbf{B}, \mathbf{T}_{g'})$$

be feature-equivalent perturbation CMMs on observed perturbations $\mathcal{A}_o$, with injective deterministic mixing maps. Let $\tau = g^{-1} \circ g'$ be the transition induced by Thm. 1. Then, for every $a \in \mathcal{A}_o$,

$$\langle W'(a-a_0), c\rangle - \frac{1}{2}\|W'(a-a_0)\|^2 = \langle W(a-a_0), \tau(c)\rangle - \frac{1}{2}\|W(a-a_0)\|^2.$$

Equivalently, after centering at c_0,

$$\langle a - a_0, (W')^\top(c - c_0)\rangle = \langle a - a_0, W^\top(\tau(c) - \tau(c_0))\rangle.$$

Under the sufficient-diversity condition of von Kügelgen et al. [2025], this is their perturbation-identifiability equation, with the remaining orthogonal-rigidity step supplied by their theorem. The same centered identity satisfies Thm. 4 with $\varphi(a) = a$, so feature equivalence extrapolates to every a_{ex} satisfying

$$a_{\text{ex}} - a_0 \in \operatorname{span}\{a - a_0 \mid a \in \mathcal{A}_o\}.$$

Proof. Substituting the Gaussian mean-shift potential into the transported identity $\Delta^{\bar{\mathbf{Q}}'}_{a,a_0}(c) = \Delta^{\bar{\mathbf{Q}}}_{a,a_0}(\tau(c))$ gives the first display. Subtracting the same identity at c_0 gives the centered display. The identifiability conclusion is the rigidity theorem of von Kügelgen et al. [2025], and the extrapolation conclusion follows from Thms. 3 and 4. □

F CMM Translations of prior work

Setting	Reference	A	Λ	C	X
Nonlinear ICA	Hyvarinen et al. [2019]	auxiliary variable $\mathbf{u}$	$(q_i(\cdot \mid \mathbf{u}))_{i=1}^k$	$\mathbf{s} \in \mathbb{R}^k$	$\mathbf{x} = f(\mathbf{s})$
	Khemakhem et al. [2020a]	conditioning variable $\mathbf{u}$	$(f(\mathbf{u}), T, Q)$	$\mathbf{z}$	$\mathbf{x} = f(\mathbf{z}) + \varepsilon$
	Khemakhem et al. [2020b]	conditioning variable by	$(\mathbf{f}_\theta, \mathbf{g}_\theta(\mathbf{y}))$	$\mathbf{x}$	$\mathbf{x}$
CRL	Squires et al. [2023]	environment $k \in \{0\} \cup [K]$	structural factor B_k	Z	$X = GZ$
	Buchholz et al. [2023]	environment $i \in I \cup \{0\}$	$\lambda_i = (B^{(i)}, \eta^{(i)}, t_i)$	Z	$X = f(Z)$
	von Kügelgen et al. [2023]	environment $e \in \mathcal{E}$	local-mechanisms $\lambda = (p_i^\lambda(c_i \mid c_{\mathrm{pa}_{\mathcal{G}_\lambda}(i)})_{i=1}^n)$	$\mathbf{V}$	$\mathbf{X}$
	Varıcı et al. [2025]	environment $a \in \mathcal{E}$	local-mechanisms $\lambda = (p_i^\lambda(c_i \mid c_{\mathrm{pa}_{\mathcal{G}_\lambda}(i)})_{i=1}^n)$	$\mathbf{Z}$	$\mathbf{X} = g(\mathbf{Z})$
Perturbation modeling	Ahuja et al. [2022b]	action a	weak-supervision mechanism (S_a, δ_a)	Z	$X = g(Z)$
	von Kügelgen et al. [2025]	perturbation label $a \in \mathbb{R}^K$	latent mean $W(a - a_0)$	$C \in \mathbb{R}^k$	$X = g(C)$
Topic modeling	Arora et al. [2012]	document d	topic mixture $w_d \in \Delta^{r-1}$	topic $T \in [r]$	word $V \in [n]$
	Huang et al. [2016]	document d	topic mixture $w_d \in \Delta^{r-1}$	topic $T \in [r]$	word $V \in [n]$
Others	Rajendran et al. [2024]	atomic concept label $\alpha = (a, b)$	$\alpha = (a, b)$	Z	$X = f(Z)$
	Schmidt et al. [2025]	group element g	$\mathbf{R}(g)$	$(\mathbf{x}, \mathbf{x}')$	$(\mathbf{y}, \mathbf{y}')$

Table 2: Variable-level CMM translations for representative prior work. Each row records the objects along the chain $A \to \Lambda \to C \to X$, following notations in the original papers. In the mechanism-based CRL rows, $\mathcal{G}_\lambda$ denotes the induced DAG from local mechanisms λ.

Setting	Reference	$\mathbf{Q} \in \mathcal{K}^{\mathrm{mk}}(\mathcal{A} \to \mathscr{L})$	$\mathbf{B} \in \mathcal{K}^{\mathrm{mk}}(\mathscr{L} \to \mathcal{C})$	$\mathbf{K} \in \mathcal{K}^{\mathrm{mk}}(\mathcal{C} \to \mathcal{X})$
Nonlinear ICA	Hyvarinen et al. [2019]	$\mathbf{T}_{a \mapsto (q_i(\cdot,a))_{i=1}^k}$	$\mathbf{B}(dc \mid \lambda) = \exp\left\{\sum_{i=1}^k \lambda_i(c_i) - \Gamma(\lambda)\right\} dc$	$\mathbf{T}_f$
	Khemakhem et al. [2020a]	$\mathbf{T}_{a \mapsto (f(a),T,Q)}$	$\mathbf{B}(dc \mid \theta, T, Q) = \frac{Q(c)\exp\{\langle \theta, T(c)\rangle\}}{Z(\theta)} dc$	$\widetilde{\mathbf{K}}\mathbf{T}_g$
	Khemakhem et al. [2020b]	$\mathbf{T}_{a \mapsto (\mathbf{g}_\theta(a), \mathbf{f}_\theta)}$	$\mathbf{B}(dc \mid \gamma, \mathbf{f}) = \frac{\exp\{-\langle \gamma, \mathbf{f}(c)\rangle\}}{Z(\gamma, \mathbf{f})} dc$	$\mathbf{T}_{\mathrm{id}}$
CRL	Squires et al. [2023]	$\mathbf{T}_{k \mapsto B_k}$	$\mathbf{B}(dc \mid B) = \mathrm{Law}(B^{-1}\epsilon)(dc)$	$\mathbf{T}_G$
	Buchholz et al. [2023]	$\mathbf{T}_{i \mapsto (B^{(i)}, \eta^{(i)}, t_i)}$	$\mathbf{B}(dc \mid B, \eta, t) = \lvert\det B\rvert (2\pi)^{-d/2} \exp\left\{-\frac{1}{2}\lVert Bc - \eta e_t\rVert^2\right\} dc$	$\mathbf{T}_f$
	von Kügelgen et al. [2023]	$\mathbf{T}_{e \mapsto \lambda_e}$	$\mathbf{B}(dc \mid \lambda) = \prod_{j=1}^n p_j^\lambda(c_j \mid c_{\mathrm{pa}_{\mathcal{G}_\lambda}(j)})\, dc$	$\mathbf{T}_f$
	Varıcı et al. [2025]	$\mathbf{T}_{a \mapsto \lambda_a}$	$\mathbf{B}(dc \mid \lambda) = \prod_{i=1}^n p_i^\lambda(c_i \mid c_{\mathrm{pa}_{\mathcal{G}_\lambda}(i)})\, dc$	$\mathbf{T}_f$
Perturbation modeling	Ahuja et al. [2022b]	$\mathbf{T}_{a \mapsto (S_a, \delta_a)}$	$\mathbf{B}(dc \mid S, \delta) = \mathrm{Law}(\text{latent transformed by } (S, \delta))(dc)$	$\mathbf{T}_g$
	von Kügelgen et al. [2025]	$\mathbf{T}_{a \mapsto W(a - a_0)}$	$\mathbf{B}(dc \mid \lambda) = \mathcal{N}(\lambda, I_k)(dc)$	$\mathbf{T}_g$
Topic modeling	Arora et al. [2012]	$\mathbf{T}_{d \mapsto w_d}$	$\mathbf{B}(T = t \mid w) = w_t$	$\mathbf{K}(V = v \mid T = t) = \Phi_{tv}$
	Huang et al. [2016]	$\mathbf{T}_{d \mapsto w_d}$	$\mathbf{B}(T = t \mid w) = w_t$	$\mathbf{K}(V = v \mid T = t) = \Phi_{tv}$
Others	Rajendran et al. [2024]	$\mathbf{T}_{(a,b) \mapsto (a,b)}$	$\mathbf{B}(dc \mid a, b) = \frac{p(c)q(\langle a, z\rangle - b)}{Z(a, b)} dc$	$\mathbf{T}_f$
	Schmidt et al. [2025]	$\mathbf{T}_{g \mapsto \mathbf{R}(g)}$	$\mathbf{B}(d(u, u') \mid \rho) = \mathrm{Law}(U, \rho U)(d(u, u'))$	$\mathbf{T}_{f \otimes 2}$

Table 3: Operator-level CMM translations for representative prior work. Each row records the three kernels in the CMM chain $A \to \Lambda \to C \to X$. For Hyvarinen et al. [2019], $\Gamma(\lambda)$ is the normalizer determined by the tuple $\lambda = (q_i)_{i=1}^k$, not an additional modulator component. For Khemakhem et al. [2020a], $\widetilde{\mathbf{K}}$ is a shared additive-noise channel, assumed injective on the relevant pushed-forward laws, as for Gaussian convolution. In the Gaussian CRL row, $\eta^{(i)}$ is the shift parameter used in that model, not an indexing map. In the mechanism-based CRL rows, λ denotes a tuple of local mechanisms and $\mathcal{G}_\lambda$ denotes the DAG induced by those mechanisms. For topic models, $\Phi \in \mathbb{R}^{r \times n}$ denotes the topic-word matrix, so $\mathbf{K}(V = v \mid T = t) = \Phi_{tv}$.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's scope as a theoretical framework for identifiability and extrapolation in concept modulation models, and the stated contributions correspond to the formal definitions, transition-based identifiability theorem, attribute-potential characterization, extrapolation criteria, and recoveries in Secs. 2 to 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Sec. 6 discusses the common-support assumption, excluded deterministic or support-changing interventions, exact population-level agreement within a specified CMM class, and the need for additional analysis to address finite-sample error and model misspecification.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Each theoretical result states its assumptions in the theorem statement or immediately preceding definitions, and proof sketches are included in the main text with complete proofs deferred to App. C.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: The paper is theoretical and does not report experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: The paper does not include experiments requiring code or data release.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper does not include training, validation, or test experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper does not include experiments, so statistical significance, confidence intervals, and error bars are not applicable.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The paper does not include computational experiments requiring specialized compute resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research is a theoretical analysis of latent-variable identifiability and extrapolation, uses no human-subject data or sensitive datasets, and does not release models or assets with apparent misuse risk.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

Justification: The work is foundational and theoretical, and it does not propose a deployable system, dataset, or model with a direct societal impact pathway requiring a broader-impact analysis.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release data, pretrained models, generative systems, scraped datasets, or other artifacts with high misuse risk.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [N/A]

Justification: The paper does not use existing datasets, codebases, pretrained models, or other external assets as part of its methodology or empirical evaluation.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce or release new datasets, code packages, pretrained models, benchmarks, or other research assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, participant studies, annotation tasks, or other research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: The paper does not involve human subjects, so IRB approval or equivalent review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not used as an important, original, or non-standard component of the core methodology.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.